Computational Methods for the Study of Face Perception

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Samuel Rivera, B.E., M.S.

Graduate Program in Electrical and Computer Engineering

* * * * *

The Ohio State University

2012

Dissertation Committee:

Aleix Martinez, Adviser

Kevin Passino

Vladimir Sloutsky

ABSTRACT

Our research develops computational approaches for studying face perception, and in particular emotion recognition. We have developed biologically inspired algorithms for representing and detecting the deformable face shape, and defined eye-tracking methodology to infer which aspects of gaze are relevant to a category learning and discrimination task. The first problem, shape detection, is addressed using two different approaches of manifold learning and probabilistic graphical models. In the manifold learning approach, nonlinear regression is used to learn the function relating the object image to the associated shape. This technique is useful because it does not require shape key points be defined at high contrast image regions, nor does it require an initial shape estimate. In addition, the manifold can be learned at extremely coarse resolutions. The probabilistic graph approach uses a graph to encode the homogenous structure of many natural objects. This approach is useful because the graph allows inference of dozens of landmarks and can be applied to a variety of object types. To understand category learning from eye tracking, we have developed tools which identify the variables of an eye tracking sequence which predict if a human subject has learned how to categorize an object. The methodology we developed was applied to both adult and infant subjects, and validated several variables previously used in categorization studies. The final portion of our work involves eye tracking

experiments to determine the dimensions of emotion recognition from faces. Our results give insight into the process of emotion recognition in adults and infants.

This is dedicated to my friends and family.

ACKNOWLEDGMENTS

I thank my advisers for their continued support and guidance. I also thank the CBCSL members and other students who collected and annotated data.

VITA

May 27, 1985 ..... Born - St. Croix, USVI

2007 ..... B.E. Electrical Engineering,
University of Delaware

2012 ..... M.S. Electrical Engineering,
The Ohio State University

2007-present ..... Graduate Research Assistant,
The Ohio State University

PUBLICATIONS

Research Publications

Rivera, S., Best, C., Yim, H., Martinez, A., Sloutsky, V., & Walther, D. "Automatic Selection of Eye Tracking Variables in Visual Categorization for Adults and Infants". In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 2240-2245). Austin, TX: Cognitive Science Society.

Samuel Rivera and Aleix M. Martinez. "Learning Deformable Shape Manifolds". *Pattern Recognition*, Vol. 45, No. 4, pp. 1792-1801, 2012.

FIELDS OF STUDY

Major Field: Electrical and Computer Engineering

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

This PhD work is concerned with face perception, and in particular the perception of emotion from expressive faces. It has been argued that facial expressions are an important part of communication, especially between parents and their infant before verbal language develops [122, 116]. Facial expressions may also provide important cues during social interaction. It is not surprising that populations affected by the autism disorder, a condition marked by impaired social skills, exhibit impaired recognition of emotion from faces [12]. The impairment may result from an improper encoding of the face, supported by a study showing that scan paths differ between autistic and healthy populations [58]. If that is the case, emotion recognition training may provide relief to these clinical populations. Given the ubiquity of facial expressions in social interactions as well as the potential benefit to clinical populations, we aim to understand and model the process by which emotions are recognized from emotional facial expressions.

Our work is divided into 4 main parts. The first part discusses the modeling of detailed face shape. We propose an approach to shape detection of deformable shapes in images via manifold learning with regression. Our method does not require shape key points be defined at high contrast image regions, nor do we need an initial

estimate of the shape. We only require sufficient representative training data and a rough initial estimate of the object position and scale. We demonstrate the method for face shape learning, and provide a comparison to nonlinear Active Appearance Model. Our method is extremely accurate, to nearly pixel precision and is capable of accurately detecting the shape of faces undergoing extreme expression changes. The technique is robust to occlusions such as glasses and gives reasonable results for extremely degraded image resolutions.

The second part introduces another approach to deformable shape detection using probabilistic graphical models. Standard detectors are typically limited to locating only a few salient landmarks such as landmarks near edges or areas of high contrast [39, 47, 35], often conveying insufficient shape information. This paper presents a novel approach to locate a dense set of salient and non-salient landmarks in images of a deformable object. We explore the fact that several object classes exhibit a homogenous structure such that each landmark position provides some information about the position of the other landmarks. In our model, the relationship between all pairs of landmarks is naturally encoded as a probabilistic graph. Dense landmark detections are then obtained with a new sampling algorithm that, given a set of candidate detections, selects the most likely positions as to maximize the probability of the graph. Our experimental results demonstrate accurate, dense landmark detections within and across different databases.

The third part details our work in eye tracking which determines variables relevant to category learning and use. We present a computational approach for the selection of eye tracking variables that distinguish learners from non-learners of visual categories. Previous methods for the selection of eye tracking variables have been ad-hoc

or hypothesis driven. In the absence of a good hypothesis, researchers are left to experiment with many alternatives. To resolve this problem, we use feature extraction and classification algorithms from machine learning to automatically identify the eye tracking variables that best correlate within sample eye tracking sequences belonging to the same category yet discriminate between categories. This approach allows us to extract the few (i.e., two to four) most diagnostic features from a pool of dozens. While previous work required the testing of a large number of hypotheses, we demonstrate how the proposed methodology yields the same result without the need to test a large number of alternative hypotheses. Instead, our method is data driven, i.e., the resulting model is obtained from the data. The proposed methodology was verified in a visual categorization task with adults and infants. We presented infants and adults with a category learning task and tracked their eye movements. We extracted an over-complete set of eye tracking variables encompassing durations, probabilities, latencies, and the order of fixations and saccadic eye movements. We identified a small set of variables that allowed us to predict category learning among adults and 6- to 8-month-old infants.

The final portion focuses on emotion recognition from faces. Previous research has shown that adults recognize emotional expression categories using information from distinct locations of the face and different spatial frequencies [111, 112, 30]. Smith *et al.* [112] used the bubbles paradigm [45] and found little overlap in the positions of the face that transmitted the 6 basic emotions. One drawback of using the Bubbles methodology is that the impoverished stimuli may force participants to use a local feature based strategy to recognize the emotions. This is problematic because research suggests faces are processed holistically [38].

An alternative approach to understanding which face areas encode emotion is to monitor attention during expression viewing by tracking eye gaze, since gaze is linked to spatial attention [34, 110]. Eisenbarth and Alpers [30] conducted an eye tracking study to investigate whether scan paths are specific to emotion. They presented faces expressing five emotion categories for 2500 ms each, and had participants rate the images on scales of positive or negative valence and emotional arousal. They found that participants spent a different proportion of looking times at discrete face areas when viewing different emotional expressions, although subjects looked at the eyes and mouth across the different emotions. A recent eye tracking study which asked participants to identify the emotion category of expressive faces also found that participants look qualitatively similarly in terms of percent looking time at the eyes, nose, and mouth across emotion category [50]. They spend the greatest proportion of looking time at the eyes, followed by the nose and mouth. The study suggests that although people look at local features that are diagnostic of the emotion category, participants probably extract a holistic representation of the face.

We hypothesize that if the gaze sequence differs systematically for different emotions, then the gaze sequence should be diagnostic of the stimuli's emotion category. To test this, we performed an eye tracking study where adults were asked to label the emotional expression category modeled in a series of still face photographs. We demonstrate that there is considerable overlap in the areas looked at when adults view emotional faces, consistent with the study by [50]. We found that while gaze is somewhat predictive of the emotion category, it is insufficient to characterize the stimuli emotion label. That was the case for gaze over the entire trial and segments

of the trial, and using the overall fixation pattern as well as the first order fixation sequence.

We conducted another eye tracking study with infants ranging from eight to 24 months old in order to characterize the change in scan path and the ability to recognize emotional categories over development. We found evidence that infants are able to generalize discrimination of emotional face categories across identities, although they looked at expressive faces qualitatively differently from adults. Overall, the results suggest that there are some emotion specific aspects of gaze, but the holistic face information is probably extracted across all emotional categories.

CHAPTER 2

DEFORMABLE SHAPE MANIFOLDS

Shape detection is an important problem in computer vision because of its utility in object recognition, classification, tracking, and segmentation among others [65, 72, 21, 80]. The general shape detection problem can be stated as follows: Given an image, can we delineate the shape of a specific object in the image? This problem becomes difficult when the shape is not rigid, but can deform, translate, rotate, change scale, or become occluded in the image. We are interested in this scenario.

We develop a new method for deformable shape detection based on manifold learning through regression. We illustrate the concept applied to face shape detection in Fig. 2.1. Consider this simplified illustration of the model where the u and v axes correspond to the face *image* space, while the z axis corresponds to the face *shape* space. Given a finite set of face image samples and their associated shape parameters, we wish to estimate the nonlinear face shape manifold so that we can interpolate the shape of new samples in the face image space. Once learned, the manifold provides a direct mapping $f(\cdot)$ from a new image $\mathbf{x} \in \mathbb{R}^p$ to the shape space $\mathbf{y} \in \mathbb{R}^d$, where p and d are the number of image features and the number of shape parameters, respectively. Hence, in contrast to most methods in shape detection and modeling which iteratively fit a model to an image until convergence, the shape estimate is given in a single step.

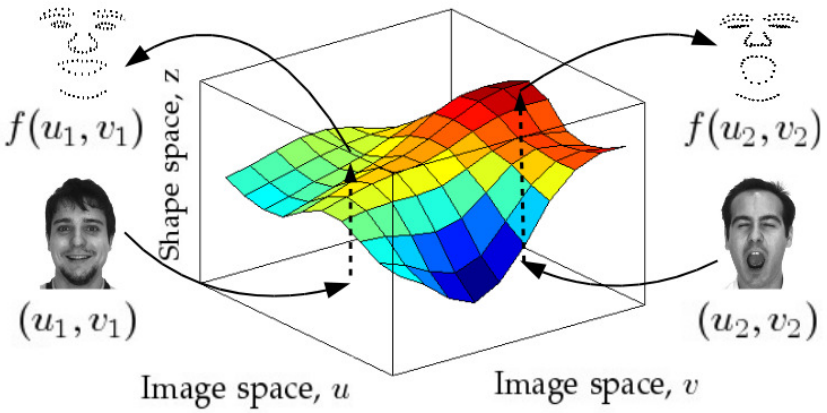


Figure 2.1: Conceptual illustration of the face shape manifold and the method presented. The u and v axes correspond to the face *image* space, while the z axis corresponds to the face *shape* space. A finite set of face image samples and their associated shape parameters are used to estimate the nonlinear manifold. This manifold defines a mapping $f(\cdot)$ from a face image sample to the associated shape parameters.

One of the first successful shape detection algorithms requiring an iterative approach was developed by Kass *et al.* [65] who utilized energy minimization to deform active contour models, called *snakes*, to fit salient image features. Thus, this method requires shapes be defined by high contrast regions. Additional constraints of how much the shape can deform (e.g., based on smoothness) are incorporated, and with a reasonable initialization of the shape, the model can deform to extract the shape of the object. A logical extension of snakes was to change the smoothness constraint of the shape for one that defines the variabilities of the object we want to model. Cootes *et al.* [21] developed on this idea with models that could only deform in ways specific to a given shape class. The shape variability was modeled using a probability density function (pdf) learned by manually delineating shapes in sample images of the object. This model is called the Active Shape Model (ASM). ASM works well

but still requires high contrast regions such as edges for fitting the active contour. To overcome this drawback, Cootes *et al.* defined the Active Appearance Model (AAM) [20], where the density of the texture is also learned from sample images. The algorithm finds the shape which best fits to the set of possible textures given by the learned pdf. This method was enhanced by using boosting to learn the shape parameter update and confidence score [71]. Other authors take a Bayesian approach to shape modeling, learning the conditional density of the shape parameters given the object image, and iterating to the maximum a posteriori (MAP) estimate of the shape parameters [48, 135, 51]. Liang *et al.* [70] improve on the idea by using regularization at accurately aligned points to reduce the occurence of local minima favored by the global shape model. Zhang *et al.* [131] further develop on this Bayesian approach by using regression to learn a sequence of unimodal conditional density functions which guide the shape estimate toward the correct solution.

Another alternative is to train a set of classifiers to detect various face fiducials and inter-connect them to estimate the shape [27]. In this case, a sliding window approach is used, where the classifiers are evaluated at all positions and scales of interest followed by pruning and voting for the final detection. This approach can provide very accurate results if high resolution images are available, but the sliding window method is computationally demanding.

Zhou and Comaniciu changed direction with Shape Regression Machine (SRM) and utilized nonlinear regression to segment the left ventricular endocardium in highly structured images [133, 134]. SRM employs boosting with an over-complete feature bank to train a strong learner which associates an image with a shape. A strong learner is a linear combination of weak learners which coorespond to the outputs

of local feature extractors. Detecting shape using such an approach is advantageous because it avoids the initialization and iteration posed by the above methods. Regression has also been used by Cristinacce and Cootes [23] to model imprecisely detected fiducials in faces. Imprecisely detected fiducials had previously been modeled using a pdf [73].

Our approach is related to SRM in that both use nonlinear regression to relate an image with a shape, but there are fundamental differences. Our model uses kernel regression with global object appearance while SRM uses boosted regression with local features. Our experiments show that the local approach is not as effective in the low resolution setting as the holistic approach. Furthermore, SRM is proposed for medical image segmentation so objects are normalized according to an estimated scale and rotation parameter. In this work, images are normalized according to the estimated eye positions which is more appropriate in the context of faces.

The approach proposed in the present paper and illustrated in Fig. 2.1 has the positive aspects of the above approaches while eliminating some of the drawbacks:

1. As the discriminative approaches, the method is non-iterative.

2. As the generative approaches, the method does not require a sliding window.

3. The method does not require strong shape contours so the shape manifold can be learned at extremely low resolutions.

4. The manifold is based on a specific shape model, so it will give a reasonable shape estimate even in the case of large occlusions, deformations, or other image changes.

In the current work, we apply the method to face shapes, but the ideas can be generalized to other deformable shapes.

The remainder of this chapter is organized as follows. In Section 2.1 we summarize regression and in particular, the methods Kernel Ridge Regression and Support Vector Regression. Section 2.2 describes our methodology in detail. Section 2.3 describes several experiments which demonstrate the method's ability to detect the shape of realistic face images with occlusions, and at extremely low resolution. We close the article with our conclusions in Section 2.4.

2.1 Regression

Regression allows us to find the functional relationship between some predictor variables $\mathbf{x} \in \mathbb{R}^p$ and an associated output $\mathbf{y} \in \mathbb{R}^d$ [105]. Given $\mathbf{x}$ and $\mathbf{y}$ from unknown distributions, we want to find the mapping function $f : \mathbf{x} \to \mathbf{y}$ which minimizes the expected risk,

$$\mathbf{E}[L(f(\mathbf{x}), \mathbf{y})],$$

where $L(f(\mathbf{x}), \mathbf{y})$ is an appropriate cost function which penalizes the deviations between $f(\mathbf{x})$ and $\mathbf{y}$. Since we do not know the underlying distribution of the independent and dependent variables, we generally minimize the empirical risk. Given a training set $(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \ldots, n$, the empirical risk is given by

$$\frac{1}{n} \sum_{i=1}^{n} L(f(\mathbf{x}_i), \mathbf{y}_i).$$

Preliminary experiments showed that a simple linear model was insufficient for representing the manifold illustrated in Fig. 2.1 accurately. We thus now turn our attention to nonlinear regression.

2.1.1 Kernel Ridge Regression

Kernel Ridge Regression (KRR) is the kernel extension of Ridge Regression (RR), which is a penalized version of linear least squares regression [59]. RR minimizes the cost function,

$$L(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{W}^T\mathbf{x}_i\|_F^2 + \lambda\|\mathbf{W}\|_F^2,$$

where λ is a user determined regularization parameter, $\|.\|_F$ denotes the Frobenius norm, and $\mathbf{W} \in \mathbb{R}^{p\times d}$.

If we create a matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ where each row is one of the vectors from the training input and a matrix $\mathbf{Y} \in \mathbb{R}^{n\times d}$ with the associated output values, the solution is given by

$$\arg\min_{\mathbf{W}} L(\mathbf{W}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2.1}$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix. Our regressed function is then $f(\mathbf{x}) = \mathbf{W}^T\mathbf{x}$, where $\mathbf{x}$ is the input vector.

It has been shown [56] that we can extend this method to the nonlinear case through the use of kernels. If we express $\mathbf{W}$ as a linear combination of the training samples $\mathbf{X}$ and set the derivative of the cost function with respect to $\mathbf{W}$ equals zero [107], the solution for the regressed function is given by

$$f(\mathbf{x}) = \mathbf{Y}^T(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\kappa(\mathbf{x}), \tag{2.2}$$

where $\mathbf{K} \in \mathbb{R}^{n\times n}$ is the Gram matrix of the training data. The entries of the Gram matrix are given by $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, where $k(\cdot,\cdot)$ a Mercer kernel [124] and $\kappa(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1),\ k(\mathbf{x}, \mathbf{x}_2),\ \ldots,\ k(\mathbf{x}, \mathbf{x}_n))^T$. We use the popular Radial Basis Function (RBF) kernel given by

$$k_\sigma(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|_2^2}{2\sigma^2}\right),$$

where σ is a parameter to tune and $\|.\|_2$ denotes the *Euclidean*-norm. We denote the σ used in KRR by σ_K. Note that multiple KRR assumes the output values, the shape parameters, are uncorrelated. Section 2.2.3 describes the shape model used to achieve this property.

2.1.2 ε-Support Vector Regression

ε-Support Vector Regression (ε-SVR) is another linear regression method which finds a linear function

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b,$$

where $\mathbf{w} \in \mathbb{R}^p$ and b is a scalar, such that the difference between the regressed output and the true output is below ε for all the training data while keeping $f(\cdot)$ as smooth as possible. Smoothness is achieved through regularization by penalizing $\|\mathbf{w}\|_2^2$ [113].

The underlying assumption of the algorithm is that there exists a function $f(\cdot)$ which can correctly predict training data with ε precision. To mitigate this assumption, one can introduce real valued scalar slack variables ξ_i and ξ_i^* to the above definition.

It has been shown [121] that this problem can be formulated as the following convex optimization problem,

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & \begin{cases} y_i - \mathbf{w}^T\mathbf{x}_i - b & \leq \quad \varepsilon + \xi_i \\ \mathbf{w}^T\mathbf{x}_i + b - y_i & \leq \quad \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq \quad 0, \end{cases} \end{aligned} \tag{2.3}$$

where $C > 0$ is a constant which controls the trade-off between the smoothness of $f(\cdot)$ and the allowable error greater than ε, and y_i is the scalar output associated

with $\mathbf{x}_i$. Instead of solving (2.3) directly, we solve the dual problem:

$$\begin{aligned} \text{maximize} \quad & \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\mathbf{x}_i^T\mathbf{x}_j \\ -\varepsilon\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i + \alpha_i^*) \end{cases} \\ \text{subject to} \quad & \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C], \end{aligned} \tag{2.4}$$

where α_i and α_i^* are the dual variables [9]. Using the dual formulation, $\mathbf{w}$ is expressed as a linear combination of training samples so that the problem is independent of the dimensionality of the input space [113]. This is important since we want to solve for $\mathbf{w}$ in the kernel space.

The dual problem can be solved using quadratic programming, yielding

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}_i, \\ f(\mathbf{x}) &= \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}_i^T\mathbf{x} + b. \end{aligned} \tag{2.5}$$

The value of b follows since it must satisfy the Karush-Kuhn-Tucker (KKT) conditions [9]:

$$\alpha_i(\varepsilon + \xi_i - y_i + \mathbf{w}^T\mathbf{x}_i + b) = 0,$$

$$\alpha_i^*(\varepsilon + \xi_i^* - y_i + \mathbf{w}^T\mathbf{x}_i - b) = 0.$$

Notice that in (2.4) and (2.5), the function input and training samples, $\mathbf{x}$ and $\mathbf{x}_i$, only appear as inner products in the original feature space. Therefore, the method can be extended to the nonlinear case through the use of a Mercer kernel, $k(\cdot, \cdot)$. The

dual problem then becomes

$$\begin{aligned} \text{maximize} \quad & \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(\mathbf{x}_i, \mathbf{x}_j) \\ -\varepsilon\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i + \alpha_i^*) \end{cases} \\ \text{subject to} \quad & \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C], \end{aligned} \tag{2.6}$$

and

$$f(\mathbf{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(\mathbf{x}_i, \mathbf{x}) + b. \tag{2.7}$$

As above, we use the RBF kernel. We denote the σ used in ε-SVR by σ_S.

ε-SVR is advantageous because it yields a sparse solution [113], but a good multiple output formulation is yet to be defined. Typically, the multiple output ε-SVR is defined by first uncorrelating the output values (shape parameters), then regressing each separately. We use this approach in our implementation along with the single output ε-SVR implementation LIBSVM [19].

2.2 Methodology

The method requires learning a function $f(\mathbf{x}) \rightarrow \mathbf{y}$ which relates an image feature vector $\mathbf{x} \in \mathbb{R}^p$ to a set of shape parameters $\mathbf{y} \in \mathbb{R}^d$ associated with shape coordinates $\mathbf{s} \in \mathbb{R}^k$. Once the shape parameters are regressed, we can obtain the associated shape coordinates by applying a mapping g from $\mathbf{y}$ to $\mathbf{s}, g(\mathbf{y}) = \mathbf{s}$. The shape parameterization and mapping functions are described in section 2.2.3. Shapes are defined by the two dimensional coordinates of 130 points (also known as *landmarks*) delineating the major facial fiducials (eyes, eyebrows, nose, mouth, and jaw) in a face image. This set of $\frac{k}{2}$ two-dimensional coordinates, where k is an even integer equal to twice the

number of shape landmarks, defines the face shape $\mathbf{s}$ for the image feature vector $\mathbf{x}$. Throughout the discussion, when we refer to the *shape* we mean the shape landmark coordinate vector $\mathbf{s}$, while *shape parameters* correspond to the associated parameters $\mathbf{y}$. The following describes the image normalization, shape parameterization, image representation, and function learning methodology in detail.

2.2.1 General Algorithm

Our first step is to normalize all faces to facilitate subsequent modeling. This involves detecting and cropping the faces, then scaling them to a standard size. Then we detect the eye positions using an approach similar to [36], where regression is used to learn the function which maps the scaled face image to the eye positions. After detection, we rotate the images to an upright view and standard inter-eye distance. Normalizing the images restricts the range of face deformations, concentrating the samples in the image space and the associated shape space.

To model imprecise eye detections in *test* faces, we normalize the *training* faces according to their perturbed eye positions following a Normal distribution. More formally, let the true eye positions for all the training data be vectorized as $true\ position = (L_x, L_y, R_x, R_y)^T$, where L and R correspond to the left and right eye, respectively, and x and y correspond to the horizontal and vertical coordinates, respectively. The perturbed positions are given by

$$perturbed\ position = true\ position + \varepsilon, \tag{2.8}$$

where $\varepsilon \sim N(\mu_e, \Sigma_e)$, and μ_e and Σ_e are the sample mean and covariance of the eye detection error. Note that we perturb the eye positions using the joint distribution

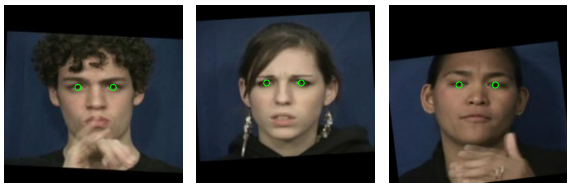


Figure 2.2: Normalized faces with automatic eye detection coordinates highlighted.

of the error since we expect the errors to be correlated. Fig. 2.2 shows some example normalized faces.

Next, we use regression to learn the face shape manifold defined by $f(\cdot)$. The input to the regressor is a cropped image region $\mathbf{x}$ centered at the mean training face position while the outputs are the associated shape parameters $\mathbf{y}$. The features used for regression can either be the pixels themselves or some pre-processing of the image. In our experiments we evaluate the pixel intensities and the C1 features of Serre *et al.* [106].

2.2.2 Feature Spaces

We experiment with two image feature spaces. Our first image representation consists of the pixel intensities which are vectorized in a raster scan fashion and normalized to unit length. The unit length normalization is done to reduce the effect of intensity by the lighting source. The second image representation uses the C1 features of Serre *et al.* [106] which are reduced in dimensionality via PCA. The C1 features are the output of the 2^{nd} layer of a $4-$layer hierarchical filter architecture modeling the hierarchy of the visual cortex. The first layer, S1, comprises of Gabor filters at various positions, scales, and orientations. The second layer, C1, comprises of *maximum* pooling operations of filter responses for the S1 layer in the same image

regions with the same orientation and within the same scale band. This pooling operation reduces sensitivity to small image perturbations. These image features are useful in our approach because these gradient based features reduce sensitivity to illumination changes and skin color.

2.2.3 Shape Modeling

Shape corresponds to the position of a discrete set of landmarks delineating the face features (eyes, eyebrows, nose, mouth, and jawline). The shape is modeled using a linear combination of a discrete set of d basis shapes usually referred to as shape modes [21]. The shape parameters are the coefficients of the shape modes. The modes correspond to directions in the shape space preserving most of the shape variance, where the shape space is defined as the span of all shape coordinate vectors. This model allows us to enforce reasonable limits on the possible shape deformation. For example, varying the parameter of the first shape mode may correspond to scaling the shape vertically. If we know the vertical range of the shape, then we could enforce a constraint on the contribution of the first mode in the final shape description. More modes can be included in the model to capture more of the possible shape variance. Following on our previous example, adding a second mode may allow us to represent vertical and horizontal shape scale changes. Representing shape using a number of modes equal to the dimensionality of the shape coordinate vector would correspond to a rotation of the shape coordinate vector in the shape space, or a change of basis.

Modes are derived using PCA. Specifically, the shape modes are defined by the primary eigenvectors of the shape coordinate covariance matrix $\mathbf{\Sigma_x} \in \mathbb{R}^{k \times k}$ for a

shape defined by $\frac{k}{2}$ landmarks in $\mathbb{R}^2$. These eigenvectors $\mathbf{p}_i, i = 1, 2, \ldots k$, are the ones associated with the largest eigenvalues, λ_i, with $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k \geq 0$.

The eigenvectors form an orthonormal basis for the shape space. This is important because this yields uncorrelated shape parameters, and recall that multiple KRR and our formulation of multiple ε-SVR assume the output variables are uncorrelated. Since most of the shape variance will be preserved by just a few of the shape modes, or eigenvectors, we can approximate shapes using a small number of parameters. The percentage of shape variance preserved by each shape mode $\mathbf{p}_i$ is given by the ratio $\frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j}$. Thus, we can keep the amount of modes that preserves a desired amount of the total shape variance.

If we arrange the principal eigenvectors into the columns of a matrix $\mathbf{P} \in \mathbb{R}^{k \times d} = (\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_d)$ with $d \leq k$, and the associated eigenvalues into a diagonal matrix $\mathbf{\Lambda} \in \mathbb{R}^{d \times d}$, the shape parameter vector $\mathbf{y}$ associated with shape coordinate vector $\mathbf{s}$ is given by

$$\mathbf{y} = \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{P}^T (\mathbf{s} - \hat{\mu}_s), \tag{2.9}$$

where $\hat{\mu}_s$ is the sample mean shape coordinate vector. Shape parameters are converted back to the original space with

$$\mathbf{s} = \mathbf{P} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{y} + \hat{\mu}_s. \tag{2.10}$$

The term $\mathbf{\Lambda}$ serves as a variance normalization which ensures that the objective function favors each shape parameter appropriately.

2.2.4 Training the Regressor

The function $f(\cdot)$ mapping $\mathbf{x}$ to $\mathbf{y}$ for KRR is obtained using (2.2), and for ε-SVR using (2.6) and (2.7). Note that KRR requires tuning the parameters λ and σ_K while

ε-SVR requires tuning the parameters C, ε, and σ_S. These parameters and the crop size for $\mathbf{x}$ are optimized using a grid search to minimize the 5-fold cross-validation error on the training set. Details are given in Section 2.3.3. For training, the face shape of the original face image is manually annotated in each of our databases. When a test image is presented to our algorithm, we detect the face position and scale using the Viola and Jones Detector [123], center and scale the image to a standard face size, then use our method to detect the face shape.

The manual shape annotations allow us to determine our final shape detection error. Error is defined by the average Euclidean distance from each landmark estimate to the corresponding manually annotated landmark over all test shapes. More formally, the error e for estimating N shapes $\mathbf{s}_j, j = 1 \dots N$ by $\hat{\mathbf{s}}_j, j = 1 \dots N$ is defined by

$$e = \frac{2}{Nk} \sum_{i=1}^{\frac{k}{2}} \sum_{j=1}^{N} \sqrt{(u_{ij} - \hat{u}_{ij})^2 + (v_{ij} - \hat{v}_{ij})^2}, \tag{2.11}$$

where u_{ij} and v_{ij} are vertical and horizontal components of the i^{th} coordinate of the shape j, and $\hat{u}_{ij}$ and $\hat{v}_{ij}$ their estimates.

2.3 Experiments

We evaluated 6 shape detection algorithms on the task of face shape detection. The algorithms consisted of KRR or ε-SVR with pixel intensities or C1 features, the nonlinear AAM of [51], and Adaboost regression with Haar fetures as in SRM [133]. The goals were to evaluate how well each algorithm: generalizes across several identities, manages occlusions and extreme expression changes, handles extreme degradation in resolution, and performs in a real world setting. The specific databases used were geared toward evaluating generalization ability, robustness to occlusions,

and performance in the real world. The training and testing partitions within a database were fixed across all algorithms to give a fair comparison.

2.3.1 Databases

The American Sign Language (ASL) database of [26] includes video sequences of 7 ASL signers. We manually annotated the face shape of $2,437$ images in the video sequence at the original resolution of 480×720 pixels. The face images in this dataset show large *variations in expression and self-occlusions* (hands can occlude facial regions when signing). Our goal with this database was to determine how well the algorithm handles these two problems.

The AR face (AR) database [74] contains frontal faces from over 100 subjects with largely varying facial expressions. We manually annotated 885 of the images at the original resolution of 576×768 pixels. Our goal with this database was to show that the learned manifold (Fig. 2.1) generalizes across several identities, i.e., it is *not subject-specific*, and can deal with extreme deformations such as a screaming face.

The Faces in the Wild (LFW) database [61] contains faces with varying facial expressions, pose changes, and occlusions at different resolutions, in different contexts, and in different photographic settings . Our goal with this database was to test the algorithm's performance in a real world setting. All faces in this database have already been detected, cropped, and scaled to a standard size of 250×250 pixels. The face coordinates of $2,610$ images were manually annotated at this scale of 250×250 pixels.

Different training percentages were used for each database according to the level of difficulty in estimating the manifold. In general, difficulty increases as the amount

of subjects and variability in the database increases. Variability corresponds to differences in illumination, pose, occlusions, and other imaging artifacts. Therefore, we used a random partition of 60% of the annotated images for training and the remaining 40% for testing in the ASL database. We used 80% for training and the remaining 20% for testing in the AR databse. The LFW database is by far the most challenging, requiring many training samples for reasonable manifold estimation. We used a random partition of 90% of the annotated images for training and the remaining 10% for testing. Since the training set was so large in the LFW experiments, we found it necessary to reduce the dimensionality of the pixel features for computational tractability in the $\varepsilon-$SVR experiments. PCA was used, where 99% of the variance was kept to preserve as much information as possible.

2.3.2 Different Resolution Analysis

To simulate the effects of low resolution we detected shape in images of different sizes. We localized all faces except those from the LFW database (already localized) in their original image using an off the shelf face detector, then cropped them in a square region much larger than the detected face following the approach of Huang *et al.* [61]. The cropped region was then scaled from 250×250 pixels to 50×50 pixels in decrements of 50 pixels. Details about the face detector and crop size are given in Section 2.3.3. All of these images were normalized to a 42 pixels inter-eye distance. 42 pixels was the mean eye distance of a random subset of images from the LFW database which were originally 250×250 pixels. To further challenge the algorithms and emphasize the performance in extremely degraded image conditions, we used images with a starting size of 250×250 pixels which were normalized to a

42, 20, 10, and 5 pixel inter-eye distance. The original annotated coordinates were scaled as necessary to serve as the face shape coordinates for each new image size. In each experiment and for each scale, we estimated the face shape over 10 trials using a different training and testing partition in each case. Example detections are shown in Figs. 2.3, 2.4, and 2.5 for the ASL, AR, and LFW databases, respectively.

Quantitative results of detection error are graphed in Fig. 2.6 as standard error of equation (2.11) in the first row, and the normalized detection error defined by,

$$e_{\text{normalized}} = \frac{e}{\text{Inter-Eye Distance}}, \tag{2.12}$$

in the second row. The normalization standardizes the error rates to account for the range of inter-eye distances. All other figures and tables report the standard error of equation (2.11). Additional results are tabulated in Table 2.1. A more detailed view of the trends can be seen in the cumulative error histogram plots of Figs. 2.7, 2.8, and 2.9.

Some noticeable trends are evident from the results in Table 2.1. First, the ASL database containing 7 subjects shows that KRR with pixel features performed best over a variety of resolutions when detecting shape over a limited range of subjects. Although the C1 features did not perform as well, it was important to evaluate their performance within the manifold learning framework because it has been argued that these features facilitate recognition of different classes of objects including shape and texture based objects, and provide invariance to illumination [106]. While pixels perform favorably for faces, other classes of objects may prefer the C1 representation.

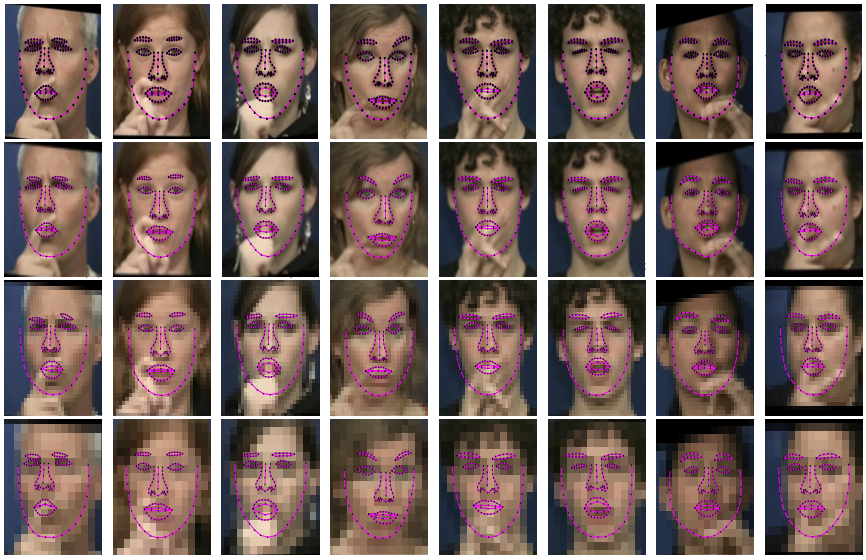


Figure 2.3: Example face shape detections for the ASL database using KRR with pixel features. Starting from the top row we display results for 250×250 pixel face images which have been normalized to a 42, 20, 10, and 5 pixel inter-eye distance.

Figure 2.4: Example face shape detections for the AR face database using KRR with pixel features. Starting from the top row we display results for 250×250 pixel face images which have been normalized to a 42, 20, 10, and 5 pixel inter-eye distance.

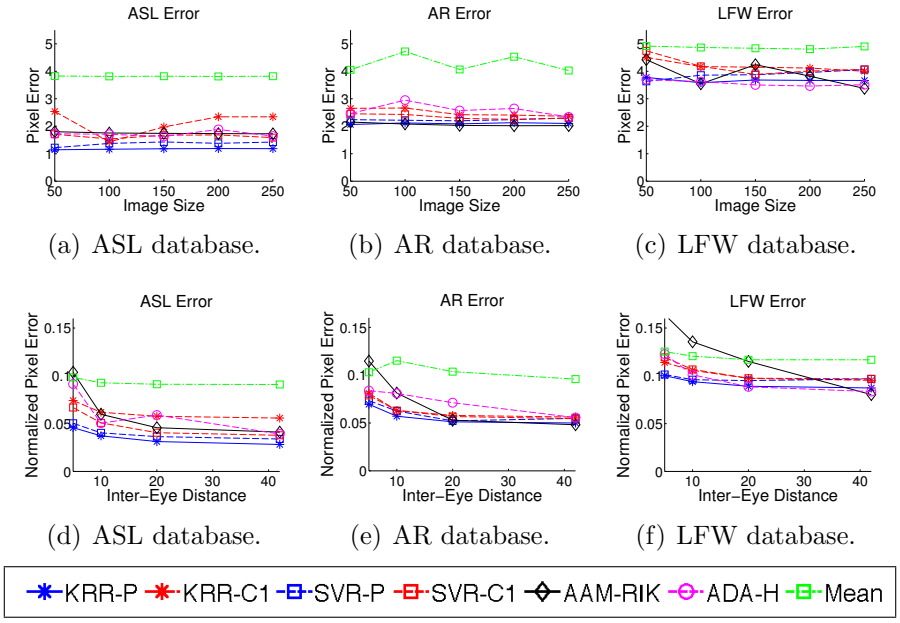
Table 2.1: Mean Euclidean landmark error of equation (2.11) and standard deviation over all test images in pixels for images at 250×250 pixels which are normalized to 42, 20, 10, and 5 pixel inter-eye distances. (a) - (c) correspond to the ASL, AR, and LFW databases, respectively. The left column entries are defined in Fig. 2.6.

		Inter-eye Distance			
		5 pixels	10 pixels	20 pixels	42 pixels
(a) ASL	KRR-P	$\mathbf{.23 \pm .15}$	$\mathbf{.37 \pm .28}$	$\mathbf{.62 \pm .53}$	$\mathbf{1.18 \pm 1.10}$
	KRR-C1	$.37 \pm .23$	$.62 \pm .42$	$1.15 \pm .81$	2.34 ± 1.68
	SVR-P	$.25 \pm .16$	$.40 \pm .31$	$.73 \pm .61$	1.42 ± 1.40
	SVR-C1	$.33 \pm .23$	$.51 \pm .42$	$.81 \pm .76$	1.59 ± 1.61
	AAM-RIK	$.52 \pm .32$	$.60 \pm .40$	$.91 \pm .66$	1.73 ± 1.30
	ADA-H	$.46 \pm .28$	$.50 \pm .32$	$1.18 \pm .80$	1.64 ± 1.12
(b) AR	KRR-P	$\mathbf{.35 \pm .21}$	$\mathbf{.57 \pm .40}$	$\mathbf{1.02 \pm .77}$	2.10 ± 1.61
	KRR-C1	$.41 \pm .28$	$.63 \pm .45$	$1.16 \pm .89$	2.36 ± 1.92
	SVR-P	$.37 \pm .36$	$.63 \pm .46$	$1.05 \pm .88$	2.30 ± 1.96
	SVR-C1	$.39 \pm .29$	$.63 \pm .48$	$1.14 \pm .92$	2.29 ± 1.89
	AAM-RIK	$.57 \pm .41$	$.81 \pm .63$	$1.06 \pm .86$	$\mathbf{2.02 \pm 1.54}$
	ADA-H	$.42 \pm .27$	$.81 \pm .59$	1.43 ± 1.02	2.34 ± 1.83
(c) LFW	KRR-P	$\mathbf{.50 \pm .34}$	$\mathbf{.94 \pm .68}$	1.78 ± 1.33	3.67 ± 2.80
	KRR-C1	$.57 \pm .37$	$1.06 \pm .74$	1.95 ± 1.37	4.00 ± 2.88
	SVR-P	$.51 \pm .35$	$.96 \pm .72$	1.90 ± 1.44	4.06 ± 3.09
	SVR-C1	$.60 \pm .40$	$1.07 \pm .75$	1.95 ± 1.41	4.07 ± 3.01
	AAM-RIK	$.82 \pm .60$	$1.36 \pm .92$	2.30 ± 2.01	$\mathbf{3.38 \pm 3.04}$
	ADA-H	$.61 \pm .39$	$1.01 \pm .67$	$\mathbf{1.77 \pm 1.29}$	3.51 ± 2.58

Figure 2.5: Example face shape detections for the LFW database using KRR with pixel features. Starting from the top row we display results for 250×250 pixel face images which have been normalized to a 42, 20, 10, and 5 pixel inter-eye distance.

The 5 and 10 pixel inter-eye distance results over all databases show that AAM-RIK and adaboost with Haar features suffer when the resolution is degraded significantly. This can be explained for the AAM-RIK by the lack of precision in synthesizing a face that is degraded to that magnitude. Similarly, the Haar features are unable to precisely describe local image cues at very low resolutions. The kernel regression based algorithms did not suffer from the degradation in resolution because the test images are not synthesized as in the AAM-RIK, and a holistic (not local) image representation is used. However, the AAM-RIK and adaboost methods yielded very accurate results at higher resolutions, as supported by the AR and LFW results in Table 2.1.

To provide a comparison to the state of the art in detailed face shape detection, the experiments were repeated for the ASL and AR database as in the work of Ding



(a) ASL database. (b) AR database. (c) LFW database.

(d) ASL database. (e) AR database. (f) LFW database.

Figure 2.6: (a) - (c) show the mean Euclidean pixel error for each database as a function of image size in pixels. Image size corresponds to the height and width of the face image before rotating and scaling to a standard inter-eye distance of 42 pixels. (d) - (f) show the normalized mean Euclidean pixel error for each database as a function of the inter-eye distance. P and C1 denote pixel and C1 features of [106], respectively. AAM-RIK denotes the AAM with Rotation Invariant Kernels of [51], ADA-H denotes the Adaboost regression [133] with Haar-like features of [123], and Mean denotes taking the mean of the training shapes as the estimate in every case.

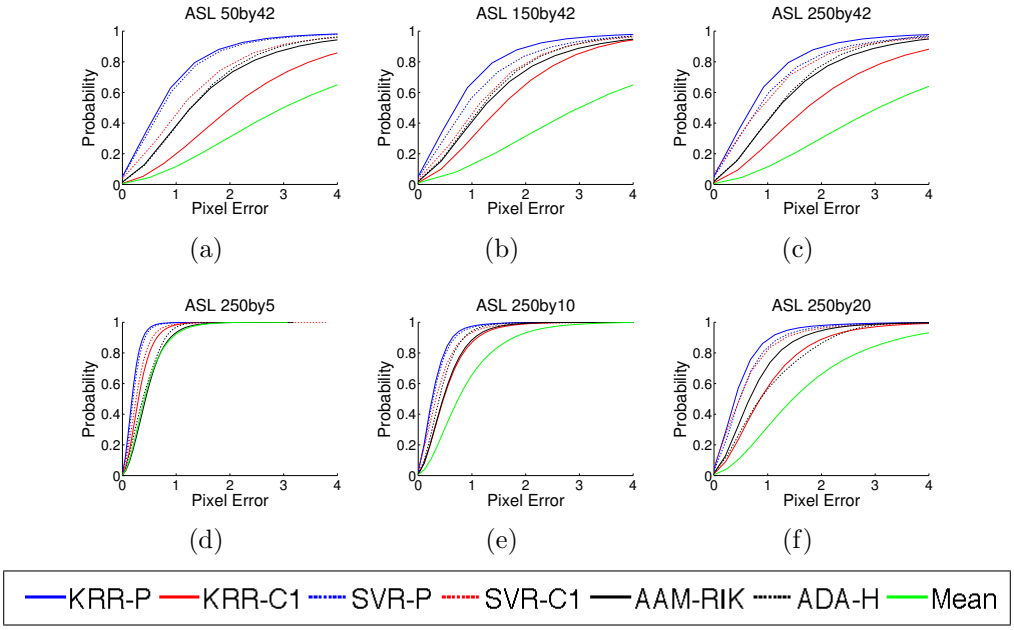


Figure 2.7: Cumulative pixel error histograms are plotted for the ASL database. The plots in (a) - (c) show error rates for images which are first scaled to 50×50, 150×150, and 250×250 pixels, then normalized to a 42 pixel inter-eye distance. (d) - (f) show error rates for images which are first scaled to 250×250 pixels, then normalized to a 5, 10, and 20 pixel inter-eye distance. Legend entries are defined in Fig. 2.6.

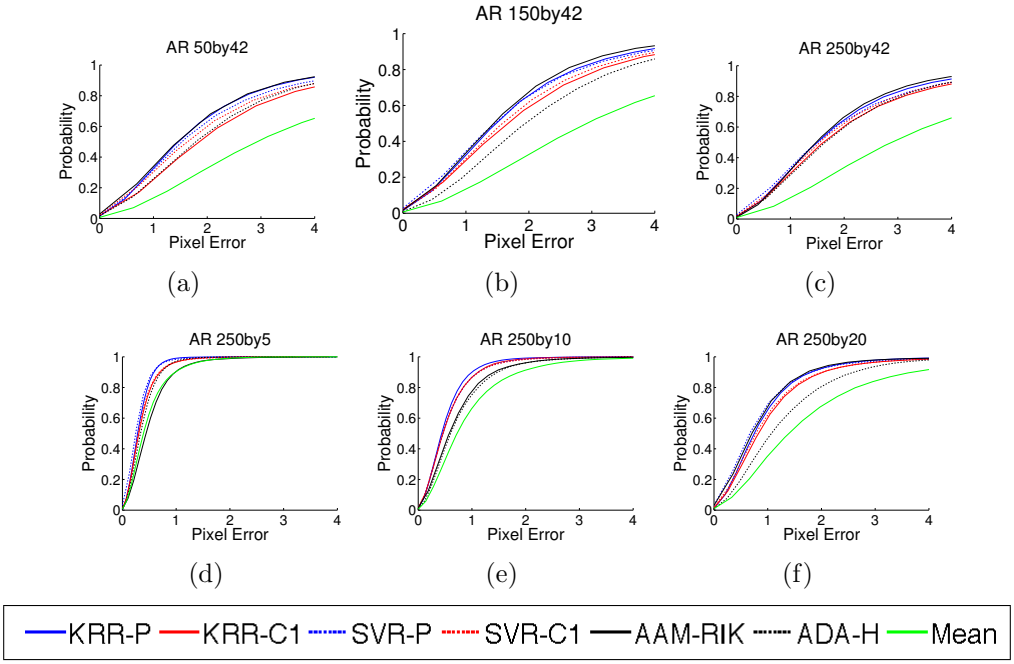


Figure 2.8: Cumulative pixel error histograms are plotted for the AR database. As above, the errors in (a) - (c) are for images scaled to 50×50, 150×150, and 250×250 pixels, then normalized to a 42 pixel inter-eye distance. (d) - (f) show error rates for images which are first scaled to 250×250 pixels, then normalized to a 5, 10, and 20 pixel inter-eye distance. Legend entries are defined in Fig. 2.6.

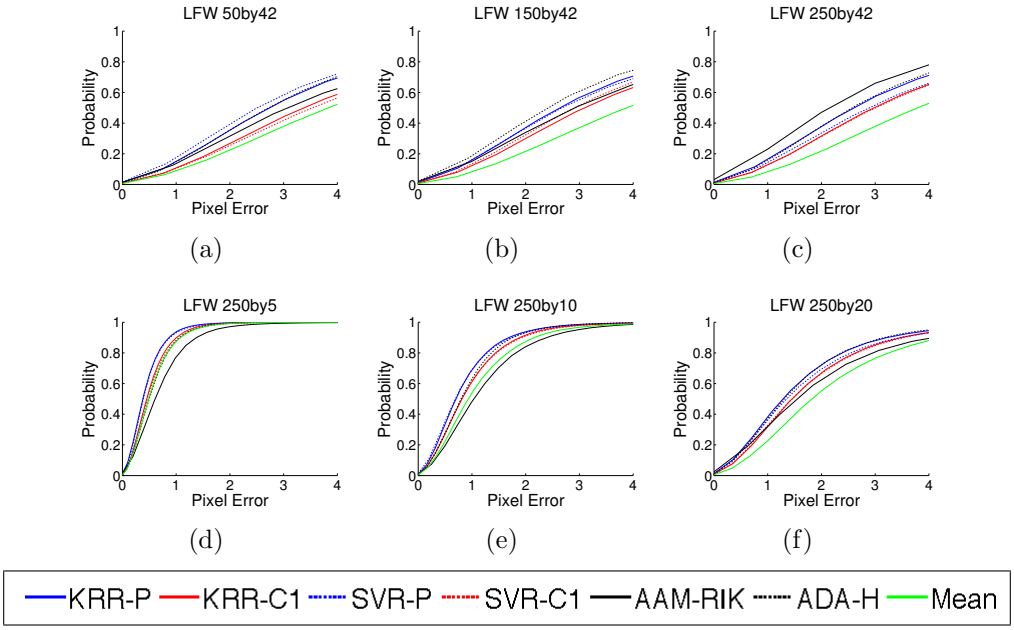


Figure 2.9: Cumulative pixel error histograms are plotted for the LFW database. The plots in (a) - (c) show error rates for images of 50×50, 150×150, and 250×250 pixels, then normalized to a 42 pixel inter-eye distance. (d) - (f) show error rates for images which are first scaled to 250×250 pixels, then normalized to a 5, 10, and 20 pixel inter-eye distance. Legend entries are defined in Fig. 2.6.

and Martinez [27]. Their algorithm relies heavily on color and edge cues, so higher resolution images are required. Therefore, we performed our eye detection in the images which were kept at the original resolution. After eye detection, the faces were scaled to an inter-eye distance of 120 pixels for the ASL database, and 111 pixels for the AR Face database, corresponding to the average inter-eye distances within the databases at the original resolution. In [27], the authors report an average pixel error of 6.9 pixels with a standard deviation of 1.5 for the ASL database, and an average pixel error of 8.4 pixels with a standard deviation of 1.2 for a combination of 400 images from the AR database and 800 images from the XM2VTS database [78]. We achieved an average error rate of 4.30 pixels with a standard deviation of 3.80 for the ASL database, and an average error rate of 6.97 pixels with a standard deviation of 5.61 for the AR database.

The error rates favor the algorithm presented, but the comparison is not easy to make for a few reasons. Namely, the algorithm in [27] is not based on regression, so subsets of the database were not sequestered for testing. Therefore, evaluation was not performed using the same images. This is especially the case for the AR results which Ding and Martinez pool with a database not used here. However, the comparison along with the others presented demonstrates that we have achieved performances above the state of the art.

To illustrate the performance with occlusions, we refer the reader to Fig. 2.3 which contains sample results from the ASL database. These are particular examples where occlusions are prominent in some of the fiducials. The unoccluded fiducials are mostly unaffected, which is expected since these fiducial shapes are regressed with the local texture information. In largely occluded regions such as the mouth in the

5^{th} column, there is a good shape guess. We can achieve this because of the shape model learned from the training samples.

To emphasize the effect of degrading the image resolution, results for each database are shown for images which are normalized to an inter-eye distance of 42, 20, 10, and 5 pixels before doing the shape detection. By comparing the results for the degraded resolutions to the original resolution, we can see that the accuracy degrades, but marginally compared to the reduction in resolution. This is not the case for the AAM-RIK and ADA-H, which break down as the inter-eye distance gets too small.

In some cases, the images with a 5 pixel inter-eye distance appear to have a large error (such as for the eye shape). However, this results from the image interpolation which produce the appearance of eyes which sag under the true eye positions. You can see this effect if you compare the 5 pixel inter-eye distance images to the 42 pixel inter-eye distance images.

To illustrate the performance in a more challenging setting, where the manifold must generalize across subjects with more extreme facial expressions, we show sample results from the AR face database in Fig. 2.4. We can see from some examples that the results are not as accurate as the previous database. This is also clear from the results in Fig. 2.6 and Table 2.1. The mouth and jawline are not aligned perfectly, but the estimate is very close.

For the realistic setting where there is little control over the pose, illumination, photographic setting, or even identity of the subject, we show results from the LFW database in Fig. 2.5. Much like the results in Fig. 2.3, the detection is not as precise as in the ASL database (zooming into the image may illustrate this better). However, the estimate is still very accurate, as seen in Fig. 2.6 and Table 2.1. Of particular

interest is the 3^{rd} column, which has very large occlusions. Much like in the ASL database, a good estimate is given.

2.3.3 Implementation Details

Faces are first detected using the openCV implementation [13] of the Viola and Jones face detector [123]. The faces are scaled to a standard size following the approach of Huang *et al.* [61]. A region 2.2 times the detected face size is cropped around the detected face position and scaled to a standard square size, then normalized as described in Section 2.2.1. We define the face shape by a set of 130 landmarks delineating the eyes, eyebrows, nose, mouth, and jaw. We use 5-fold cross-validation with a grid search on the training data to tune the crop size for $\mathbf{x}$, and the regression parameters. We choose the parameters which minimize the mean squared error of estimating the shape parameters y over the 5 validation sets. The parameters are obtained in the first experimental trial, and used for the remaining trials. The joint distribution of the eye position detection error is then estimated using the estimation errors from the cross-validation trial corresponding to the optimal parameters.

To simplify the computation of the proposed approach we proceed as follows. Shape modes are stored for faster function evaluation since they only depend on the training data. In addition, the $\kappa(\mathbf{x})$ in the KRR solution given in (2.2) is arranged into a matrix for all training samples, allowing batch shape detection in all images. Our MATLAB implementation which utilizes this batch procedure and stores the learned shape modes can detect shapes at .084 seconds per image, or 11.9 frames per second on a 2.4 GHz Intel Core 2 Duo with 4 GB RAM. This timing includes the eye detection and image rotation, and was conducted on the ASL dataset with

150×150 pixel images which are normalized to a 20 pixel inter-eye distance. 60% of the images were used for training, with the detection being performed on the remaining 40%. Faster running times can be achieved by storing the inverted matrix and Gram matrix of the KRR solution since they only depend on the training data.

$\varepsilon-$SVR yields a sparser model than KRR which can also be stored for fast function evaluation. Unfortunately, it requires we train a separate model for each shape mode.

The nonlinear AAM implementations is based on [51] which employs nonlinear shape and texture models through the use of Rotation Invariant Kernels (RIK) [52]. The model is trained separately for the different scales and resolution, using the training image subsets which are normalized to the size and resolution of the testing set. The kernel parameters are selected as suggested by the authors. In testing, we first detect and normalize the images using the approach described in Section 2.2.1, where KRR with pixel features is employed since it consistently yields accurate results. Then, we initialize the AAM with 10 random training face shapes which are aligned in position and scale with the test image, and update the shape estimate until convergence or a maximum of 30 iterations. Careful checks were made to prevent divergence. The final shape estimate is given by the converged appearance model that best approximates the face texture.

The Adaboost implementation is based on the work of [133], where piecewise functions of single Haar feature outputs are selected in a greedy manner to build a strong learner for an image based regression task. At each round of boosting, the weak learner which most reduces the residual error is added to the strong learner. Shrinkage is used to prevent over-fitting. Specifically, each weak learner is scaled by a real positive value less than 1 before being added to the strong learner. The strong

learner has several parameters: the type and number of weak learners to use, and the shrinkage value. We choose a strong learner comprising of 400 weak learners with a shrinkage factor of .1. Piecewise linear functions are employed as weak learners. For a given Haar feature, the piecewise functions are split into evenly spaced bins by 40 knots according to the maximum range of that Haar feature output from the training data. Each piecewise linear function for each bin is obtained by fitting the output from a single Haar feature to the residual error using (2.1) with regularization parameter $\lambda = .1/n$.

The training and testing images are first normalized to an upright view and standard scale using the approach described in Section 2.2.1, where KRR with pixel features is employed since it consistently yields accurate results. Then, a strong learner is learned which maps these normalized face images to their associated shape modes. Similarly to the $\varepsilon-$SVR implementation, the strong learner is an ensemble of single output strong learners for each shape mode. Shape modes are selected to preserve 90% of the shape variance, as described in Section 2.2.3.

2.4 Conclusion

We have presented a new algorithm for deformable shape detection that is based on manifold learning through nonlinear regression. By taking this approach, the algorithm has the benefits of generative and discriminative methods while avoiding the major drawbacks associated with generative, energy minimization based, and sliding window methods. Our algorithm is non-iterative and does not require shapes be defined by salient edges. Therefore, as we have demonstrated experimentally, the algorithm can be made to work at extremely coarse resolutions. Additionally, we learn

a shape model based on training data which ensures a reasonable shape estimate even in the case of large occlusions.

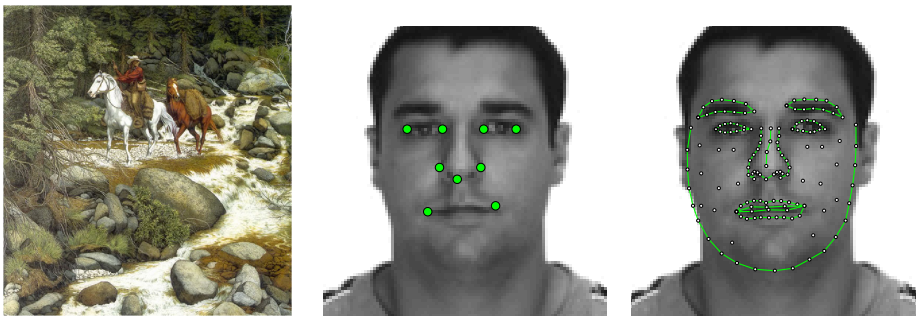
Although we presented certain feature spaces and regression techniques, these are only particular choices for a general shape detection framework based on manifold learning with supervised training data. The learned manifold relates the image feature space to the corresponding shape space. What is important about the feature space used is that it preserves the local image information. As we have demonstrated, the framework presented is flexible enough to model deformable faces, and can be trained reliably given the constraints on the training data.

CHAPTER 3

PROBABILISTIC GRAPHS FOR DENSE SHAPE DETECTION

As discussed in the previous chapter, shape detection is an important problem in computer vision whose the goal is to accurately detect the $2D$ position of specific shape landmarks, or *fiducials*, in an image. Some applications such as 3D reconstruction and the recognition of facial expressions require that the deformable shape be described by a *dense* set of *salient and non-salient* landmarks for satisfactory results. Unfortunately, current detection algorithms are typically tailored to locating only a few salient landmarks [39, 47, 35]. Some exceptions are the 3-Dimensional Morphable Model (3DMM) [11] and the $3D$ model of [49] which find a dense set of face landmarks. However, these methods require a $3D$ database to construct the model and, thus, cannot be learned directly from an image collection.

Here we propose a novel algorithm to accurately detect a dense set of salient and non-salient landmarks in an image. It does not require $3D$ object databases and can be used to design landmark detectors for different types of objects – *eg*, faces, hands, and structures in medical images. Our approach utilizes the fact that many object classes exhibit a homogeneous structure such that any detected landmark provides contextual information that facilitates the detection of the other landmarks.



(a) From salient to non-salient features (b) Standard detectors (c) Our dense detection

Figure 3.1: Can you find the 10 faces in (a)? These faces are difficult to see until a face feature is detected (*eg*, a nose); then the entire face becomes salient. (b) The output of a standard face landmark detector is typically restricted to a few salient points. (c) Our novel method provides dense detections that include both salient and non-salient landmarks.

For instance, Fig.3.1(a) shows a classical example with human faces as objects of interest. While at first sight one may not perceive the 10 faces in this image, once a few fiducials have been detected (*eg*, an eye or the nose), the remaining facial parts become readily apparent. Thus, the location of a fiducial provides information on where to find the others.

Our method hinges on the value of context - each fiducial provides information about the others. By combining all the sources of information, we can form a consensus about the whole shape. This approach of combining different contextual sources of information has been a central theme of computer vision since the late 1970s when research was focused on scene interpretation: determining the class and $3D$ position of objects producing an image. The computer vision pioneers understood that using context appropriately is crucial to solving the problem.

The development of ACRONYM [14] which was later refined by Brooks [15] demonstrated that very noisy low level features such as edges could be used to interpret a scene as long as the features were considered in the context of a detailed $3D$ object model. Real objects were modeled as generalized cones with specific geometric and algebraic constraints that were used to join and filter the coarse image features in a meaningful way. In the same vein, Hanson and Riseman developed the VISIONS system [54] for scene understanding which used domain specific knowledge to reason about low level features in an image. Barrow and Tenenbaum's Intrinsic Images [7] were also concerned with understanding the inherent $3D$ structures producing an image, but relied more on constraints associated with the physical world and image acquisition than detailed high level models.

A decade later, Strat and Fischler developed CONDOR [117] which interprets scenes by using simple features with a dictionary of *context sets* to interpret natural scenes. Context sets consist of information that is known about the world, relationships between items in the world, and how to proceed given certain conditions are met in an image. The algorithm sequentially invokes and refines the active context sets until a consistent image interpretation is found. Although the algorithms are quite different, they all emphasize the value of context in scene interpretation through high level models and physical constraints associated with the real world and image acquisition.

Our method imposes such constraints through a probabilistic graphical model. The relationship between every pair of landmark positions is encoded by the edges a graph, where each node represents a landmark position and its local texture. The local texture information of salient landmarks allow them to be detected reliably, whereas

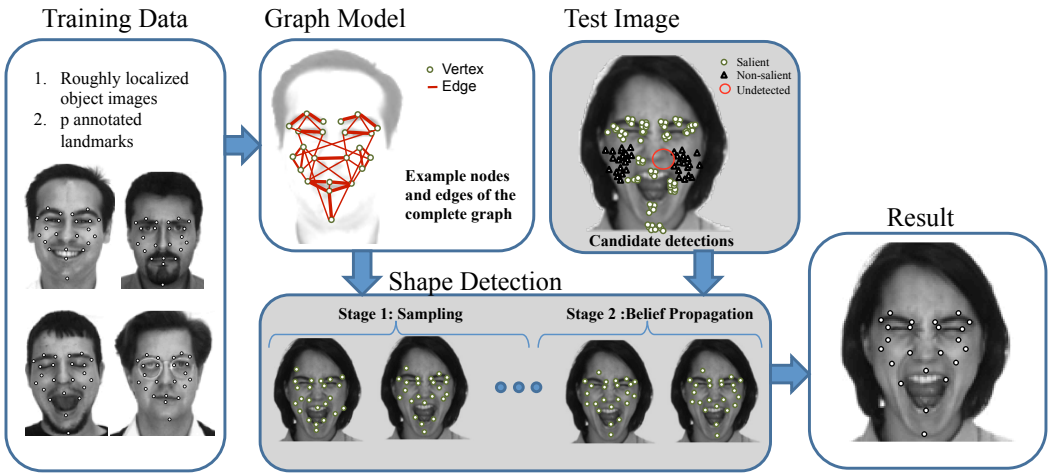


Figure 3.2: Illustration of proposed methodology: The 'Graph Model' block shows the graph learning phase of the method, where we model the relationship between salient and non-salient landmarks as a probabilistic graph. Thicker edges represent larger weights between fiducials. Only a subset of edges of the fully connected graph are shown. The 'Test Image' block highlights cases of misdetections and multiple detections for a particular fiducial. The 'Shape Detection' block shows that the graph model is used along with the local feature detections to determine the most probable shape configuration.

non-salient landmark detection is unreliable from just the local texture. Fortunately, the reliable detections constrain the position of non-salient landmarks and vice versa. In addition, the coarsely localized non-salient fiducials aid estimation of other non-salient and misdetected landmarks. As a key result, our detection algorithm can robustly estimate the positions of fiducials in areas such as face cheeks, where a simple local feature detector would generally fail, Fig.3.1(b)-(c). Hence, the resulting algorithm can be used to estimate dense landmark maps in 2D images as in 3DMMs, without requiring prior 3D models.

Our proposed methodology is depicted in Figure 3.2. Using our graphical model, landmark detection amounts to maximizing the joint probability of the graph's nodes in an image. To accomplish this, we propose a sampling algorithm which selects the most likely fiducial positions given a set of candidate detections. This algorithm can deal with missing detections, false positives, and occlusions. In addition, we show how to augment the graph and use the original low-level detections to infer many more landmark coordinates in an incremental fashion. Our experimental results demonstrate accurate, dense landmark detections within and across different databases.

3.1 Related Work

Since the 1970s computer vision has been concerned with the recognition of a particular person's face versus another [6, 55], due to the social significance of faces. These recognition algorithms, similar to the current methods, compare a standardized probe face image to a database of face candidates to find the best match. Standardization typically required precise detection of the face position and scale, and many algorithms were developed to do this automatically. They can be broadly categorized as knowledge based, feature based, template based, and appearance based [127]. See Yang *et al.* [127] for a survey.

Knowledge based algorithms are top-down methods that define a set of rules for identifying the face from clutter based on knowledge about the face, such as symmetry. These methods require a hand crafted set of rules for locating the face and removing false detections. For example, Yang and Huang [126] use rules obtained from low-resolution, or mosaic face images that preserve the essential face structure.

An example rule is that the center part of the mosaic face has a uniform color, corresponding to the blurring of the skin color at the center of the face. Edge information is used to remove incorrect detections. Kotropoulos and Pitas [67] also construct rules using mosaic images. The drawback of knowledge based methods is defining a set of rules that reliably discriminate faces from non-face patterns.

The feature based methods, on the other had, are bottom-up approaches that assume there are some distinctive features that facilitate detection. Edges, for example, have been used to detect the face outline [109]. Another approach is to detect salient fiducials such as eyes and nostrils using a low level feature detector, and determine the most plausible configuration of all fiducials using random graph matching [68]. Other researchers have relied on skin or hair texture in the images [18, 5, 60]. The drawback of these methods is that they do not deal well with occlusions or other image artifacts that disrupt the face appearance.

Template based methods rely on a characteristic face pattern called a template. The templates are predefined by experts, and the existence of a face in the image is determined based on the correlation of the template with a an image patch. Templates have been defined in terms of edges [98], silhouettes [100], and relative intensities [101] to name a few.

The appearance based methods take the idea of templates a step further by learning the template from many example faces, instead of relying on an expert to define the template. The benefit of appearance methods versus the standard template approaches is that the appearance methods automatically discover the relevant statistics of the face that distinguish it from non-faces. The appearance approaches typically define an appearance model in terms of a pdf learned from many face examples, or a

discriminative function learned from face and non-face examplex. For example, the Active Appearance Models (AAM) [20] and 3DMM [11] use a probabilistic shape and texture model learned from a set of annotated training samples to detect faces in images. The other approach is to train a classifier to discriminate between face and non-face objects, and scan the entire image for faces. Boosted classifiers have become especially popular because of their real time performance [123].

The algorithm presented in this chapter belongs to the appearance based methods and is concerned with the precise detection of a dense set of landmarks around the face. Previous methods such as AAM and 3DMM are able to detect fiducial points, but such models cannot detect variations beyond what is specified in the training set. Furthermore, the global shape model often favors a configuration similar to the mean shape that fails to capture subtle important changes such as eye blinks or single eyebrow motion. On the other hand, algorithms that rely on fiducial detections are able to fit salient key points reliably without being overconstrained by a global model. However, these approaches can yield unrealistic shape estimates when the global shape is not constrained. These methods have advanced considerably in recent years, with algorithms that even rival human manual annotations [27, 81, 132, 69, 118]. However, they provide a very limited number of fiducials around salient features such as the eyes, nose, and mouth, and most require high-quality images.

Our new method overcomes the above shortcomings by utilizing the positive aspects of fiducial detection and probabilistic shape and texture model approaches. We take advantage of advances on local feature detection to ensure that subtle shape changes are not missed by our method. Each landmark position takes into account the position of *all* other locally-detected fiducials to generate a plausible configuration

for the whole set of detected points. If some landmarks cannot be detected or are misdetected by the local feature detector, the other fiducials will constrain estimation of their positions.

Graphical models have previously been used to guide the fiducial position estimation, although using different approaches which are limited to a very small number of landmarks. Felzenszwalb and Huttenlocher [39] use a tree structured graph to infer the location of 5 face fiducials. For tree based graphs, a poorly localized root node negatively influences all daughter nodes. In contrast, our model represents a dense interconnection of landmarks. Every node has influence from all other shape landmarks so the effect of a few poorly localized nodes is circumvented by information from other nodes in the graph.

In another work, Everingham *et al.* [35] model the joint probability of 9 fiducial positions in faces using a mixture of Gaussian trees. Fiducial are found using a discriminative model with Haar-like features. This algorithm only detects stable points (salient features) and the graph is solely used for robustness to different poses. Gu and Kanade [47] use local feature detections to generate a set of candidate positions for each fiducial, and then select the most probable set of fiducials using a Bayesian model which encodes the object pose and shape. Their algorithm is initialized using an AAM and a set of fiducials are localized around each landmark. Our algorithm also relies on a set of local feature detections but the highly inter-connected structure of our graph allows us to infer positions of non-salient features while being more robust to occlusions. In general the proposed method outperforms these probabilistic graphical models because it can infer potentially dozens of fiducial positions including non-salient ones. In addition, it easily extends to non-face objects.

3.2 Probabilistic Graphical Model

Many natural objects are highly structured. Human faces, for example, exhibit strong relationships between the positions of different parts of the face; one can estimate the right eye position reliably given the position of the left eye and the nose by symmetry. In this and other objects, the information from each detected fiducial can be used to infer all other landmarks including those which are unseen, misdetected, or poorly localized. Given this insight, an appropriate modeling scheme describes the pairwise relationship between all fiducials in an object as well as the global configuration. We model the affinity between two fiducials i and j using a potential function Φ_{ij}, and the global configuration using the potential function β. β ensures that the global configuration is reasonable. The logical way to combine these sources of information is using a probabilistic graph.

We define the joint probability of p fiducial *positions* as,

$$P_{pos}(X) = \beta(X)\frac{1}{Z_{pos}}\prod_{i=1,j=1}^{i=p,j<i}\Phi_{ij}(x_i, x_j), \tag{3.1}$$

where X encodes the set of coordinates $x_1, ..., x_p$, $x_i \in \mathbb{R}^2$ has the $2D$ coordinate of fiducial i, Z_{pos} is called the *partition function* which makes (3.1) behave as a probability density function (pdf), and $\beta(X) = \exp\left(-\frac{\alpha}{2}(\hat{X}-\mu)^T\Sigma^{-1}(\hat{X}-\mu)\right)$ is the Mahalanobis distance from the translated and scale invariant shape $\hat{X}$ to the mean shape, μ. To obtain $\hat{X}$, we centered the shape to the origin and then normalize by its Frobenius norm. $\mu = \frac{1}{N}\sum\hat{X}_i$ and $\Sigma = \frac{1}{N}\sum(\hat{X}_i-\mu)(\hat{X}_i-\mu)^T$ are the mean and the covariance matrix of the training samples. The parameter $\alpha \in [0, 1]$ controls the penalty of differing from the mean shape.

Assuming that the object is detected and scaled to a standard size, the displacement between two fiducials can be modeled as a bivariate normal distribution. Although displacement between different pairs of fiducial may vary in scale, the correlations may be the same. Therefore, it is important to use a normalized distance, such as the Mahalanobis distance, when measuring the displacement. Thus, the *potential functions* $\Phi_{ij}(\cdot,\cdot)$ are defined as

$$\Phi_{ij}(x_i, x_j) = \exp\left(-(1-\alpha)\bar{w}_{ij}(\Delta_{ij} - \mu_{ij})^T \Sigma_{ij}^{-1}(\Delta_{ij} - \mu_{ij})\right), \tag{3.2}$$

where $\bar{w}_{ij}$ is the weight of the edge between fiducials x_i and x_j, $\Delta_{ij} = (x_i - x_j) \in \mathbb{R}^2$ is the $2D$ pairwise distances between landmarks i and j for a particular shape, and the parameters μ_{ij} and Σ_{ij} are the sample mean and sample covariance of Δ_{ij} which are estimated from the training data. The relationship between landmark positions is encoded as pairwise distances to make the model translation invariant. Since the quadratic term in $\Phi_{ij}(\cdot,\cdot)$ is a Mahalanobis distance, it is maximum when $\Delta_{ij} = \mu_{ij}$, and monotonically decreases from there.

In the case of face shapes, the edge connecting the eye to the eyebrow should have a larger weight than the edge connecting the eye to the mouth because the eye more strongly constraints the position of the eyebrow than the mouth. Therefore, we scale the Mahalanobis distances by normalized positive scalar edge weights $\bar{w}_{ij} \in [0, 1]$, to account for this type of variation. The normalized edge weights are defined as $\bar{w}_{ij} = \frac{w_{ij}}{\sum_{k=1}^{k=p}\sum_{l=1}^{l<k} w_{kl}}$, where the edge weights w_{ij} specify the relative importance of the edges in the graph.

Specifically, large w_{ij} means the relationship between fiducials ij will be emphasized relative to pairs with smaller w_{ij} in (3.2). Because the graph encodes the pairwise

structure of the shape in order to constrain the low level detections, an edge connecting nodes ij should have a large weight if knowing the position of node i strongly constrains the position of node j. The degree to which i constrains j is specified by the sum of the eigenvalues of the covariance matrix Σ_{ij} because the eigenvalues describe the variance along the principal axes of the joint distribution $N(\mu_{ij}, \Sigma_{ij})$. A larger variance means we know less about the relative positions of fiducials; the distance between connected fiducials can take on a very wide range of values. However, a smaller variance means the relative position of fiducial j can be inferred with more certainty. Therefore, we set $w_{ij} = \frac{1}{\|\Sigma_{ij}\|_F}$ by noting that $\|\Sigma_{ij}\|_F$ equals the square root of the sum of the eigenvalues of Σ_{ij}.

The graph described by equation (3.1) is sufficient to constrain shape estimates to be reasonable, but it is also important to consider the local texture. This is achieved by weighing the pdf by the probability that each landmark is correctly localized, where each texture implies a probability of correct localization. This texture weighting implies that the textures for each fiducial are assumed to be conditionally independent. The graph describing the joint pdf of the landmark *positions and textures* is defined as,

$$P(X) = \frac{1}{Z}\beta(X) \prod_{i=1,j=1}^{i=p,j<i} \Phi_{ij}(x_i, x_j) \prod_{k=1}^{p} \gamma_k(x_k), \tag{3.3}$$

where $\gamma_i(\cdot)$ is a potential function of the i^{th} fiducial's local texture. More formally, these functions γ_i are the normalized confidences of the local detection. We calculate that confidence using a kernel based density estimation algorithm [46], and normalize such that $\gamma(\cdot)$ behave as a probability.

3.3 Testing Procedure

3.3.1 Fiducial Detection

Fiducial position estimation relies on a set of local detections. The local detections are a result of evaluating a classifier for each fiducial at each image patch within the image regions expected to contain the associated fiducial. The regions are determined by the training data. The classifiers rely on features extracted from the image patches, and are based on the local texture or *context* features. Salient points like corners of the eyes prefer local texture (pixel value features) as in [27], while non-salient points such as points in the cheeks require a more global texture feature [108]. The classifiers are learned using Kernel Linear Discriminant Analysis (KLDA) [79, 8] and the annotated training data. KLDA finds a low-dimensional projection of the image features $\mathbf{x} \in \mathbb{R}^d$ which maximally separates samples from different classes while grouping members of the same class. We use the Radial Basis Function (RBF) kernel defined as $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right)$, where σ is a parameter to select, and $\mathbf{x}_i$ are the vectorized sample images. The kernel parameter is optimized using the criterion of [128].

Each local detector returns a set of candidate positions $D_i = \{\hat{x}_i^1, ..., \hat{x}_i^{m_i}\}$, where $\hat{x}_i^j$ is the j^{th} candidate position of fiducial i, and m_i is the number of candidate detections for fiducial i. These candidate detections consisting of their position and corresponding texture provide information about the unknown true fiducial positions in the graph. The main task of this model is to find the set of candidate detections that maximize the graph probability, given their position and local texture. More formally, the objective is

$$X^* = \arg\max_{x_1,...,x_p} P(X), \quad s.t. \;\; x_i \in D_i, i = 1 \ldots p. \tag{3.4}$$

3.3.2 Estimation of the Fiducials

Estimating fiducial positions amounts to optimizing the objective function of equation (3.4). The size of our search space depends on the number of candidates per fiducial. This space is very large and evaluating the probability of each configuration requires $\prod_{i=1}^{p} m_i$ trials, which is computationally expensive. To overcome this problem, we first focus our search on a reduced set of highly probable configurations. The most likely configuration initializes the second stage, a belief propagation algorithm, that provides the final shape estimate.

To find a set of likely fiducial configurations, the optimization algorithm assumes that at each iteration, $p-1$ "known" fiducials (salient and non salient) are coarsely detected. We then randomly sample for the p^{th} fiducial knowing that its pdf is conditioned by the "known" ones. This process is iterated for a different set of fiducials until a likely global configuration is found. More formally, to find the $\hat{x}_i^{j_i}, i = 1 \dots p$, which maximizes (3.3), we initialize x_i randomly to one of the local detections in $D_i, i = 1 \dots p$. In the *first stage*, we sequentially update $x_k, k = 1 \dots p$, by taking a random sample from the conditional pdf of x_k given all other nodes:

$$P(x_k|x_{-k}) \propto \gamma_k(x_k) \exp\left(-\frac{1-\alpha}{n} \sum_{j \neq k} \bar{w}_{kj} (\Delta_{kj} - \mu_{kj})^T \Sigma_{kj}^{-1} (\Delta_{kj} - \mu_{kj}) \right.$$
$$\left. + \frac{\alpha}{2} x_k^T \Sigma^{-kk} \hat{x}_k \right), \tag{3.5}$$

where $x_{-k} = \{x_1, x_2, \dots, x_{k-1}, x_{k+1} \dots x_p\}$, $\hat{x}$ is the k^{th} fiducial of $\hat{X}$, $\Delta_{kj} = x_k - x_j$ and Σ^{-ji} is the inverse of the submatrix that measures the covariance of the i^{th} and j^{th} fiducial of scale and translation invariant shape $\hat{X}$. This procedure is repeated for all the fiducials x_k to generate a sample $\tilde{X}_i$, where $i = 1, \cdots, maxIter$. Note

that the conditional pdf of equation (3.5) can be calculated explicitly and normalized for the set of discrete candidate positions. We store the final configuration and the corresponding probability from equation (3.3) after every iteration.

To be robust to empty and false positive candidates resulting from occluded or noisy images, we use the probabilistic graph of equation (3.1) to augment the sets D_i with the maximum-likelihood (ML) estimate of each fiducial position x_i given x_{-i}. Taking the logarithm of equation (3.1) yields a quadratic function. Taking derivatives, we find that

$$x_k^{MLE} = ((1-\alpha)G_p + \alpha G_s)^{-1}((1-\alpha)W_p + \alpha W_s) \tag{3.6}$$

where

$$G_p = \left(\sum_{j\neq k} \bar{w}_{kj}\Sigma_{kj}^{-1}\right)^{-1},\ G_s = s_c\Sigma^{-kk}, s_c = \|X\|^2$$

$$W_p = \left(\sum_{j\neq k} \left[\bar{w}_{kj}\Sigma_{kj}^{-1}(x_j - \mu_{kj})\right]\right),\ W_s = s_c\left(\Sigma^{-kk}\mu_k + \sum_{j\neq k}\Sigma^{-kj}(\hat{x}_k - \mu_k)\right).$$

μ_k is the k^{th} coordinate of $\mu + t_x$, where t_x is the centroid of X. Note that this procedure without taking into account x_k^{MLE} is commonly know as a Gibbs Sampling [44].

In the *second phase* of the optimization, we initialize with the previously sampled configuration $\tilde{X}_i$ that maximizes equation (3.3), and repeat the sequential procedure until convergence or a maximum number of iterations. However, instead of drawing random samples from the conditional pdf, we select the value $x_k \in D_k$ which maximizes equation (3.3). The procedure is summarized in algorithm 1.

Algorithm 1 Inference Algorithm

Input=$\{D_1, ..., D_p\}$
for $i = 1$ to p **do**
  Set x_i^0= random sample of $\tilde{D}_i$
end for
for *stage* = *sampling* to *beliefpropagation* **do**
  for $t = 1$ to *maxiter* **do**
    for $i = 1$ to p **do**
      Set $\tilde{D}_i = \{D_i, x_i^{MLE}\}$
      Calculate $P(x_i^{t-1}|x_{-i}^{t-1})$ using (3.5) and the candidates $\tilde{D}_i$
      if *stage* = *sampling* **then**
        Let x_i^t be a random sample of $P(x_i|x_{-i})$
      else
        Let x_i^t be $\arg\max_{\hat{x}_i^j \in \tilde{D}_i} P(x_i^{t-1}|x_{-i}^{t-1})$
      end if
      Let $x_i^{t-1} \rightarrow x_i^t$
    end for
  end for
  Set x^0 as the x^i that maximize (3.3)
end for

3.4 Experiments

3.4.1 Face landmark detection

We first evaluate the proposed algorithm on three face databases: AR [74], Labeled Faces in the Wild (LFW) [61], and the XM2VTS database [77]. Faces are roughly localized in position and scale using the Viola-Jones face detector, then cropped and scaled to 150×150 pixels, corresponding to a mean inter-eye distance of approximately 15 pixels. For the AR database, we train with 448 images, and test on 448 images containing subjects not found in the training set. For the LFW database, we train with 1027 images, and test on 500 images. For the XM2VTS database, we train with 448 images, and test on 350 images containing subjects not found in the training set. We also use the model trained with the LFW database to detect fiducials on the AR and XM2VTS databases. The error is measured as the Euclidean distance from the ground truth fiducial positions (*ie*, manual markings) to the estimated position for a

Figure 3.3: We show example results of the derived approach. From top to bottom, the rows correspond to the shape detections for the AR database [74], the LFW database [61], and the XM2VTS database [77]. The database contains face images in unconstrained environments. Results show robustness to occlusions, pose, and lighting.

total of 50 fiducials over all test images. Results are compared to those of the AAM algorithm.

Example detections using our algorithm are shown in Fig. 3.3 with error histograms shown in Fig. 3.4. The first three histograms are for within database testing. That is, the training and testing set, although disjoint, are from the same database. The last three histograms are for across database experiments. In this case, we employed the training set of one database and do the testing on the images of the other specified database. These are the most challenging and interesting experiments. Figure 3.5 shows an additional visual comparison between our method, [20] and [35].

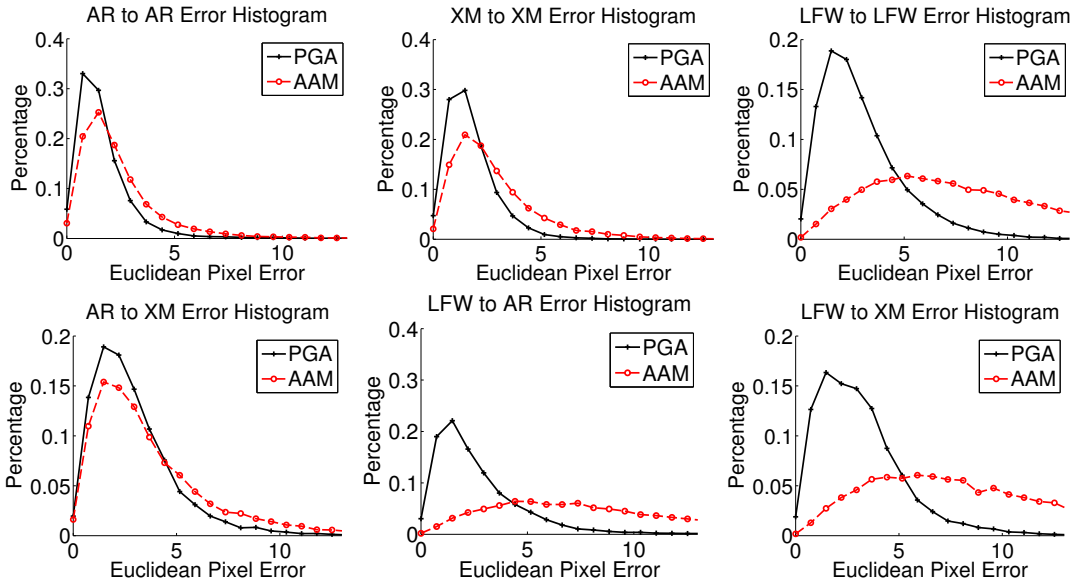
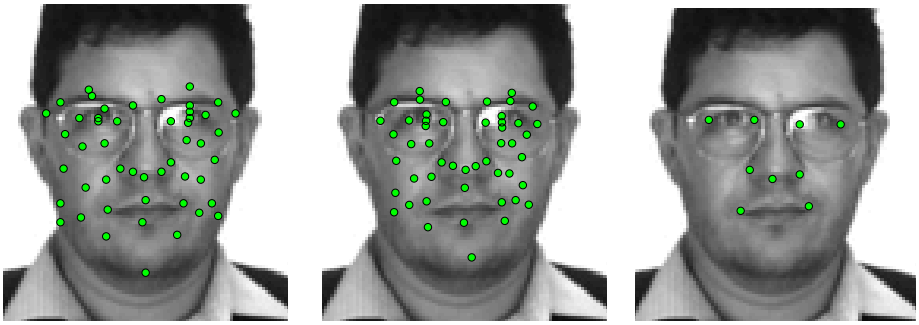


Figure 3.4: Error histograms (Euclidean distance, in pixels) for a total of 50 detected fiducial points *versus* the ground truth of the testing sets. The ground truth positions were obtained by manual annotation. PGA denotes our new probabilistic graph algorithm, while AAM denotes the classical AAM.



(a) Our Method (b) AAM (c) Result of [35]

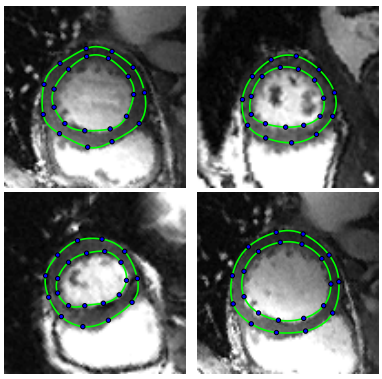
Figure 3.5: Comparison of our algorithm with AAM [20] and the local detector of [35]. Our method provides more precise detections than the AAM, and many more fiducial points than the local feature detector.

Our algorithm outperforms the AAM in every case. For the case of constrained databases such as the XM2VTS and AR databases, our algorithm slightly outperforms the AAM. However, for the case of the unconstrained LFW database, and the across-database experiments, our algorithm performs significantly better than the AAM. This reiterates the problem of learning a global probabilistic shape and texture model. It also verifies the benefit of using a probabilistic graphical combined with local detection of fiducials.

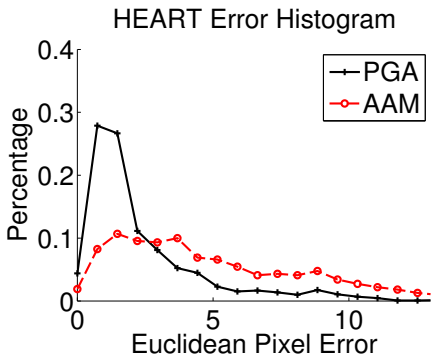
We also compare our method to the algorithm of [35] on the AR and LFW databases for the 9 fiducial points detected by their algorithm. The implementation was obtained from the authors' website, and was already trained. On the AR database, our algorithm achieves an error with mean and standard deviation of 1.5315 ± 1.2751 pixels while [35] achieves an error rate of 1.9189 ± 1.4619 pixels. On the LFW database, our algorithm achieves an error with mean and standard deviation of 2.5052 ± 1.6972 pixels, while [35] obtains an error rate of 2.2801 ± 4.1774 pixels. On the XM2VTS database, our algorithm achieves an error with mean and standard deviation of 1.5561 ± 1.7515 pixels, while [35] obtains an error rate of $1.7922 + -1.8497$ pixels.In summary our method provides more accurate detections and additional set of non-salient landmarks.

3.4.2 Cardiac MRI

To demonstrate the flexibility of our method, we now show detections of landmarks describing the epicardial and endocardial contours of the left-ventricle (LV) on cardiac magnetic resonance images (MRI). Our image database includes images of 8 subjects with a total of 160 images that were rescaled to 100×100 pixels. For each subject,



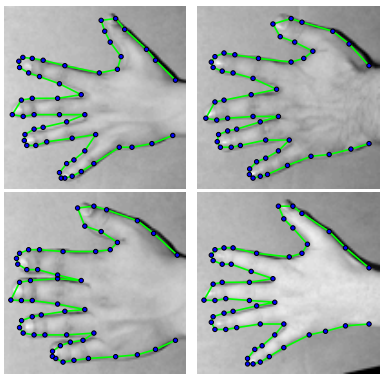
(a)



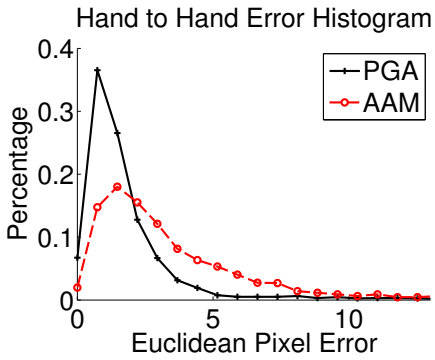
(b)

Figure 3.6: Heart Experiments. (a) Detected landmarks delineating the epicardial and endocardial contours of the LV in cardiac MRI (best seen in color). (b) Error histogram (Euclidean distances, in pixels) for 22 detected landmarks *versus* the ground truth of the testing images.

20 images depict the contraction and relaxation of the LV in short-axis view during a complete cardiac cycle. Each image was manually marked by a cardiologist with 22 landmarks around the LV. We randomly split the data into 100 images for training and 60 images for testing. Fig. 3.6(a) shows examples of detected landmarks and the approximate LV boundaries for 4 different subjects. Note that the detected contours correctly segment papillary muscles and trabeculae together with the interior blood pool of the LV. The boundaries were obtained by interpolating the Fourier transform of sequences of landmark coordinates represented as complex numbers, $l_i = x_i + y_i\sqrt{-1}$. Following the same comparison procedure of Section 3.4.1, our algorithm outperforms AAM with a mean error and standard deviation of 2.4049 ± 2.2949 pixels while AAM achieves an error rate of 5.0019 ± 3.5206 pixels. The error histograms are shown in Fig. 3.6(b).



(a)



(b)

Figure 3.7: Hand Experiments on [115]. (a) Detected landmarks delineating the shape of the Hand. (b) Error histogram (Euclidean distances, in pixels) for 52 detected landmarks *versus* the ground truth of the testing images.

3.4.3 Hand Shapes

To further challenge the algorithm, we detected 52 landmarks describing the hand contour. The image database of [115] contains 40 images of 4 subjects showing different hand shapes. We rescaled the images to 100×100 pixels, then for 3 folds of cross validation, we randomly selected 30 images for training and 10 images for testing. Fig. 3.7(a) shows examples of the detected landmarks and the approximate contour of the hand.

Following the same comparison procedure of Section 3.4.1 and combining the 30 testing results of all cross validation folds, our algorithm outperforms AAM with a mean error and standard deviation of 1.8805 ± 2.3707 pixels while AAM achieves an error rate of 3.5008 ± 3.0955 pixels. The error histograms are shown in Fig. 3.7(b).

3.4.4 Incremental Learning: Inferring Additional Landmarks

So far we have shown two cases where the algorithm can be used for detecting shapes having salient and non-salient landmarks. This is sufficient for most applications, but some require a shape representation with many more landmarks. Given the assumption that a set of localized fiducials constrains the position of other shape landmarks, we should be able to accurately predict the position of $m >> p$ fiducials based on a localized subset of p fiducials obtained using the above approach. To achieve this, note that equation (3.6) only relies on the shape part of the graphical model (no reliance on texture). Therefore, we can learn the parameters for an m node graph from an annotated set of images, and simply estimate the position of each of the undetected fiducials as the ML estimate of that fiducial position given the position of the p previously detected fiducials using equation (3.6). Fig. 3.8 shows the much denser shape detection achieved using the 50 fiducial detections of Fig. 3.3 and an augmented shape model. Note that no local feature detection or iterative sampling was done to infer the much denser set of landmarks.

3.5 Conclusion

We have presented a new algorithm for detecting a dense set of salient and non-salient landmarks in images of a deformable object. This method exploits the fact that each landmark position constrains the position of other landmarks on an object. In our model, the pairwise relationships between landmarks is naturally encoded in the nodes and edges of a probabilistic graph. Given a set of candidate landmarks provided by local detectors, our algorithm selects the set of candidate detections that maximize the joint probability of the graph. Experimental results including training

Figure 3.8: We show a denser set of fiducial positions which were inferred using the detections from Fig. 3.3 and an expanded probabilistic graph as in Section 3.4.4.

and testing across different datasets show that our method outperforms specialized, salient feature detectors as well as AAMs. Besides being more accurate, note that our method also provides dense detections and is not restricted to faces.

CHAPTER 4

IDENTIFYING RELEVANT CATEGORY LEARNING VARIABLES

Human behavior can be strikingly complex, making the inferences about the underlying cognitive and neural processes a daunting task. One way of alleviating the problem is to use methods that, due to high sampling rates, produce large volumes of data per subject per trial. For example, eye tracking while meeting this criterion, is easy to use and can be employed with a wide range of human participants. Furthermore, eye movements are closely linked to attention and use the same neural circuitry as shifts of attention [22]. This tight link between eye movements and attention as well as its high temporal fidelity make eye tracking an excellent tool for studying phenomena such as visual attention, categorization, category learning, and many others. This leaves us with the task of identifying those aspects of the rich stream of data produced by modern eye trackers that are best suited for answering a given scientific question. Variables as varied as saccade latency, fixation frequency or density, or dwell time on a particular area of interest (AOI) have been used in the existing literature. Here we introduce a new principled method that allows us to identify the most useful eye tracking variables fully automatically. We demonstrate

our method with an object categorization task. We first verify the method with adult subjects and then apply it to data from infants.

Categorization is the process of forming an equivalence class, such that discriminable entities elicit a common representation and/or a common response. While category learning exhibits early onset [92], relatively little is known about the underlying mechanism and the development of early categorization. The primary reason is the limited duration of infants' cooperation, yielding only a small number of data points per participant. These limitations have restricted researchers' ability to answer fundamental questions about categorization in infants: How do infants learn a category? And does this process undergo development?

Analyses of eye movements may help solve some of these problems. Eye movements are tightly linked to visual attention (see Rayner, 1998, for a review) and they yield multiple (albeit not necessarily independent) data points even for relatively short trial durations. Therefore, analyses of eye movements can provide critical information of how attention allocation changes in the course of category learning. However, eye tracking results in a large amount of data, and it is not clear a priori what (if any) components of eye movement are related to category learning. As a result, eye tracking researchers are free to choose from a large set of variables without a common set of principles for deciding which or how many variables to analyze. In the following, we review the infant category learning eye tracking literature in order to substantiate this claim. Given the limited number of eye tracking categorization studies with infants, we take a broader approach and review studies that examine categorization, object completion, and visual attention.

One variable that has been used across a variety of tasks is saccade latency [3, 62]. For example, Johnson, Amso, and Slemmer (2003) examined whether learning affects object representations in infancy. Four- and six-month-old infants were presented with an object that moved behind an occluder and then reemerged on the other side of the occluder. The researchers reasoned that if babies maintain the existence of the occluded object, they should anticipate the object to reemerge from the occluder. In this case, participants should exhibit a faster eye movement to the point of reemergence than if they do not anticipate the object. In two other studies [2, 4], researchers used saccade latency to examine the development of visual selection. Participants (3-, 6-, and 9-month-olds and adults) were presented with a Spatial Negative Priming (SNP) paradigm. On a given trial they were shown an attention grabbing target in Location 1 and a discreet distracter in Location 2. On the next trial, they were either shown the target in Location 2 (a negative priming probe) or in Location 3 (a control trial). SNP was inferred from greater saccade latency on the probe trials than on the control trials.

Other potentially informative variables are (a) frequencies of fixations per unit of time within one or more Areas of Interest (AOI), (b) dwell times within one or more AOIs, and (c) frequencies of saccades within and between AOIs [64]. In one study, Johnson and colleagues (2004) examined the development of object unity perception in infancy using both behavioral and eye tracking data. In the task, participants were habituated to a rod moving behind an occluder. After participants habituated, the occluder was removed to reveal either a broken rod or a complete rod. Infants who perceived the rod as moving behind the occluder as a coherent object, indicated by a recovery of looking after habituation, were identified as perceivers. Participants who

perceived a broken rod did not recover their looking after habituation and were considered non-perceivers. The authors then examined eye tracking data for perceivers and non-perceivers using the eye tracking variables described above.

Perhaps the most frequently used eye tracking variable is fixation location. Researchers have relied on this variable across a variety of tasks, including object completion [3, 63], understanding other people's actions [37], and a variety of categorization and category learning tasks [10, 76, 93]. For example, Quinn et al. (2009) examined categorization of cats and dogs in 6- to 7-month-olds. They found that when items were presented in the canonical upright position, categorization accuracy was associated with a high proportion of looking to the head; whereas, when items were presented in an inverted position, categorization was associated with the large proportion of looking to the body. Best et al (2010) presented 16- to 24-month-olds with a category learning task. Categories included artificial items that had shapes in four locations, with two of the shapes being category relevant (i.e., present in all members of the category, but not in non-members) and two being irrelevant (i.e., exhibiting both within- and between-category variability). The researchers examined the proportion of fixations to category-relevant features and its change in the course of familiarization. A summary of the reviewed studies is presented in Table 4.1.

Source	**Task**	**Eye Tracking Variable**
Johnson, et al., PNAS, 2003	Object completion	Saccade latency
Amso and Johnson, Developmental Psychology, 2006	Object completion	Proportion fixation to AOI
Johnson, et al, Infancy, 2004	Object completion	Fixation and saccade frequency, dwell time
Johnson, et al, Developmental Psychology, 2008	Object completion	Proportion fixation to AOI
Falck-Ytter, et al, Nature Neuroscience 2006	Goal perception	Proportion fixation to AOI, AOI fixation time
Amso and Johnson, Cognition, 2005	Visual Search	Saccade latency
Amso and Johnson, Infancy, 2008	Visual Search	Saccade latency
Quinn et al, Child Development, 2009	Categorization	Proportion fixation to AOI
McMurray and Aslin, Infancy, 2004	Category Learning	Proportion fixation to AOI
Best, Robinson, and Sloutsky, Proc. of CSS, 2010	Category learning	Proportion fixation to AOI

Table 4.1: Comparison of previous eye tracking variables.

Two conclusions can be drawn from this review. First, multiple eye tracking variables have been used across studies to examine infants' learning. And second, although all these variables make intuitive sense, no formal selection process of these variables has been defined. This poses several concerns and questions. Namely, since different variables are used in different studies, do these variables correlate and thus provide redundant information? If not, why should any one variable be used instead of another? Should the variables be selected based on the specific categorization task, or should a fixed subset of eye tracking variables be used across all studies? Can we define a principled way of determining which variables to analyze in a given category learning experiment? This chapter defines a methodology to address these questions and concerns.

Our approach works as follows. We extracted a large set of possible variables from the adult and infant gaze sequence during a categorization task (e.g. fixations, saccades, gaze sequences, etc). Some of the variables have been used in analyzing categorization experiments, whereas others were new. Our goal is to use the power of statistics and machine learning to identify eye tracking variables that best predict category learning in adults and subsequently in infants.

The *significant contribution* of this work is that it provides a systematic methodology for identifying eye tracking variables that are linked to category learning, thus allowing researchers to better understand category learning from eye tracking data. Furthermore, our results retrospectively validate the use of several variables in the eye tracking studies mentioned above.

4.1 Methods

4.1.1 Participants

This project was approved by a Behavioral and Social Sciences IRB at The Ohio State University. Adults were given sufficient information and signed a consent form before freely participating in the study. At least one parent of each infant participant provided written consent before the study.

Three category learning experiments were conducted, two focused on adults and one on infants. Twenty-four adults participated in Experiment 1. Forty-six adults who did not participate in Experiment 1 participated in Experiment 2. All adult participants had normal or corrected to normal vision and were undergraduate students at The Ohio State University participating for course credit. In Experiment 3, fifteen 6- to 8-month-old infants participated in the experiment. All parents reported their infants to be developing typically and without known health problems.

4.1.2 Materials

Category members were flower-like objects with six petals. An example object is shown in Fig. 4.1, with the petals enumerated for clarity. Those numbers were not shown during the experiment. There were four different categories, each defined by a single petal having a distinguishing color and shape. Specifically, the category defining features were category A: a pink triangle at position 4; category B: a blue semi-circle at position 4; category C: an orange square at position 6; and category D: a yellow pentagon at position 6. Each object was uniquely associated with one category. That is, no one object exhibited the defining features for two or more categories. Stimuli were displayed on the computer subtending approximately 11×11 degrees of visual

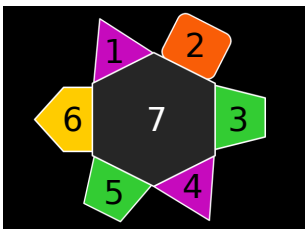


Figure 4.1: An example category object with the Areas of interest (AOI)s enumerated. Numbers were not displayed to the participants. The category defining features were category A: a pink triangle at position 4; category B: a blue semi-circle at position 4; category C: an orange square at position 6; and category D: a yellow pentagon at position 6.

angle.The eccentricity of the stimuli subtended an approximate horizontal visual angle of 14.4° and an approximate vertical visual angle of 11.5°.

During all three experiments, the participants' eye gaze was recorded using a Tobii T60 eye-tracker (Falls Church, VA) at the sampling rate of 60Hz while participants sat approximately 60 cm away from the display screen.

4.1.3 Experiment 1 - Adult supervised learning

To validate the efficacy of the approach before applying it to infants, adult participants were tested. Prior to learning in Experiment 1, adults were instructed that there was a single feature defining a category. This hint introduces supervision and previous research indicates that supervision has consequences with respect to how quickly participants learn to classify the objects, especially when there are few overlapping features [66].

The experiment had 8 blocks. In each block there were 8 learning trials followed by 4 testing trials. In a learning trial, a category member was displayed in the center

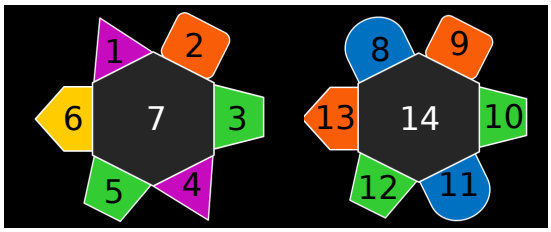


Figure 4.2: Illustration of category pair image with AOIs labeled. Numbers were not shown to participants. The relevant AOIs 4 and 6 on the left object correspond to AOIs 11 and 13 on the right.

of the screen one at a time for 1.5 seconds each. In a testing trial, a novel category member of the to-be-learned category and a novel member of a contrasting category were presented side by side, each centered approximately 7.7° horizontal visual angles from the center of the screen. An example with labeled AOIs is shown in Fig. 4.2. Test stimuli were displayed on the screen until the participant made a decision via key press about which stimulus was a member of the learned category. The left/right position of the test stimuli was counter-balanced. A randomly located fixation point (cross-hair) directed the participants gaze to a position on the monitor in-between trials. The to-be-learned category remained the same for the first 4 blocks. A second to-be-learned category was introduced in the final 4 blocks without notice to the participant. If the experiment started with a category defined by the petal at position 4 (category A or B), the second category was defined by the petal at position 6 (category C or D), and vice-versa. Using categories with definitive features at different positions provided a mechanism to verify the reproducibility of the variables determined most important.

4.1.4 Experiment 2 - Adult unsupervised learning

The procedure in Experiment 2 (unsupervised condition) was identical to that in Experiment 1 except that participants did not receive supervision (i.e., no hint provided) about the category structure.

4.1.5 Experiment 3 - Infant supervised learning

The infant experiment was conceptually similar to Experiment 1, but was methodologically adapted for infants by using a familiarization paradigm. In a familiarization paradigm, the infant is repeatedly exposed to examples of a particular category before a discrimination task where the preference for a familiarized category member and a novel category member are measured.

To aid infant learning, identical pairs of category exemplars were shown on each trial. This was also done so that the presentation of stimuli in the learning and testing phases had an identical layout. Furthermore, there was only a supervised condition, in which the infants were presented with a pre-trial fixation video of synchronized sound and motion (e.g., looming flower petal with corresponding whistle sound) to draw their attention to the single category-relevant feature. It should be noted that no unsupervised condition was conducted with infants, because previous developmental research suggests supervision is necessary for young children to learn categories with a sparse category structure [66].

Once the infant looked at the fixation video, the learning trial commenced. Infants had to accumulate 3 seconds of looking to the category exemplar pairs. Whenever an infant looked away, an attention-grabbing fixation was presented until the infant reconnected with the images on the screen. After accumulating 3 seconds of looking

to the stimulus pair, the supervisory fixation video was again presented followed by another learning image pair. This procedure was repeated for 8 blocks with 8 learning pairs per block.

In the testing phase, a novel category member was paired with a novel non-category member as in the adult experiments. The standard assumption is that an infant can discriminate between the category and non-category objects if he or she displays a novelty or familiarity preference. There were two test trials per block, in which a novel exemplar from the learned category was paired with a novel exemplar from a novel category. Test trials were presented for a fixed duration of 6 seconds, and left/right position of familiar or novel category objects was counterbalanced.

4.1.6 Collecting and filtering eye tracking data

Eye movements were monitored during object viewing with the Tobii T60 eye tracker. The system tracks eye movements by illuminating the eye with infrared light and capturing corneal reflection at a frequency of 60 Hz (i.e., every 16.6 ms). As the eye moves, the angle between the pupil and the corneal reflection increases, allowing the x-y coordinates of the gaze position to be measured over time.

Unfortunately, the gaze data contain noise, missing data, and micro-saccades, which makes identifying true fixations and saccades difficult. Therefore, we processed these data using MATLAB-based software created in our laboratory by the first author. The raw eye tracking data from every experimental block were filtered using a Kalman filter [82] before extracting the variables of interest. The eye gaze data from both the left and right eye were filtered separately. The average of the

filtered data from left and right eyes yielded the mean eye gaze data, which were used in the current analyses.

4.1.7 Labeling the Data

The eye movement sequences during the *learning phase* of the experiment aid in understanding category learning, while the sequences during the *testing phase* aid our understanding of category use. Before applying our methodology to understand these processes, however, the eye tracking data from both the learning and testing phases of the experiments were labeled as *learner* (class 1), *non-learner* (class 0), or *indeterminate* (class 2). Indeterminate samples were not analyzed.

Adult Labels: Intuitively, labels for adult data are readily identified based on the accuracy of the responses during the testing phase. An uninterrupted string of correct responses during the testing phase suggests that the participant has learned the category. Each adult experimental block yielded 12 eye movement sequences. These correspond to eye movements during the presentation of 8 exemplar images during the learning phase and 4 test images during the testing phase. Adult participants had 4 blocks of learning and discriminating the same category before switching to a new category. This amounted to 32 samples of the learning phase, and 16 samples of the testing phase for each category per participant. The 16 samples from the testing phase were associated with a 16 digit binary string, called the *response string*. This data structure shows performance over the first and last 4 blocks of the experiment. A one identifies a correct response, while a zero denotes an incorrect response on the associated test trial. An example is shown in Fig. 4.3. We labeled each 16 digit response string separately as follows.

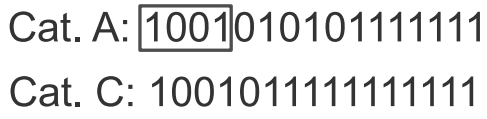


Figure 4.3: Illustration of a time series for one subject. Ones encode correct category discrimination, while zeros encode incorrect responses. The first row shows the accuracy over the first four blocks (presentation of first category), while the second row shows accuracy over the last four blocks (presentation of second category). The class labels (learner or non-learner) are determined separately for each row, because the category condition is different for each row.

We expect a learner's response string to contain a series of ones beginning within the string and terminating at the end of the response string. This pattern indicates that at some point the participant learned the category and correctly discriminated the category from that point on. A participant who has not learned the category (non-learner) would select one of the two stimuli by chance in each trial. A non-learner could get lucky and achieve a series of correct guesses. In order to determine if a participant is a learner or a non-learner we need to establish a criterion that allows us to reject chance as the cause for a series of ones. The question that we need to answer is how many ones we should expect for a learner. We address this problem by assessing how likely it is that we see a sequence of M consecutive ones in a binary response string of length $R = 16$. Under the null hypothesis, the participant does not know the category label and selects one of the stimuli by chance, giving her a 50% chance of correctly guessing the category member. Each sequence is equally likely given this assumption, so the probability of guessing at least M right in a row is the total number of sequences having M ones in a row ($(R - M + 1) \times 2^{(R-M)}$)

divided by the total number of binary sequences of length R (2^R). This yields the probability $p = (R - M + 1)/(2^M)$. For $R = 16$, $M = 10$ is the minimum number that achieves a significance level of $p < 0.01$ ($p = 0.0068$). Therefore, we rejected the null hypothesis that a participant was guessing randomly when we identified a consecutive string of 10 correct responses.

We call the position of the first correct response in this string of correct responses the point of learning (POL). The *test phase* and *learning phase* samples before the POL were labeled as non-learner, while the samples after the POL were labeled as learner. The learning phase samples from the block associated with the POL were labeled as indeterminate, because it was unclear at exactly which trial during the block the category was learned.

If the learning criterion was not achieved, we then identified the remaining non-learning and indeterminate samples. We first labeled correct responses at the end of the respond string as indeterminate. Those samples did not meet the learning criterion, but might be attributed to learning late in the experiment. The remaining samples were labeled as non-learner. Approximately 8% of the adult eye track samples were labeled indeterminate.

Infant Labels: Obviously, infants are not able to respond by keyboard to identify a category object. Instead, we used a variant of the preferential looking paradigm to determine if an infant could discriminate between novel exemplars of a familiar category object and a novel category object. Recall that the preferential looking paradigm assumes that infants who consistently look more to one class of stimuli when shown two classes of stimuli are able to discriminate between the two classes. This means that if the infant consistently looks longer at the learned category object

(or novel category object), then he or she is assumed to be discriminating between the familiar and novel categories.

Given this paradigm, we labeled each infant's gaze data by blocks. Each block consisted of two test phase samples. We determined novelty preference as the ratio of total looking time to the novel category object compared to the total looking time to the novel category plus the familiar category object. We sorted the mean of the novelty preference for each block according to the absolute difference from 0.5. A third of the blocks with mean novelty preference closest to 0.5 were labeled as non-learner. The third of the blocks with novelty preference furthest from 0.5 in absolute value were labeled learner. The samples in the middle third were labeled as indeterminate.

4.1.8 Variable List

We compiled an over-complete list of eye tracking variables. We began with the fundamental variables, fixations and saccades. Fixations occur when eye gaze is maintained at a single position for at least 100ms. They were identified using the dispersion threshold algorithm of [99]. Saccades are rapid eye movements that move the eye gaze between points of fixation. To be considered a saccade, the eye movement needed to exceed smooth pursuit velocity of 30° per second or 0.5° per sample at 60Hz [114]. The fixations and saccades were determined with respect to a specific AOI within an object. AOIs are regions of an object image or scene that can be grouped in some meaningful way, such as color uniformity or the structural nature of the object. The AOIs can further be described as relevant or non-relevant, based on their role in determining object category membership.

These fundamental eye tracking variables were combined in various ways to derive a larger set of variables. Our variable list is defined as follows:

1. *AOI fixation percentage* describes the percentage of time fixated at the different AOIs during a trial. All non-AOI fixations were discarded in this and all of the variables defined. For an image with q AOIs, this variable was encoded as a q-dimensional feature vector with a value for each AOI. The fixation percentages were normalized so that they sum to 1, unless there were no fixations at AOIs. In that case, all percentages were set to 0.

2. *Relevant AOI fixation density* is a scalar value between zero and one which describes the percentage of the total time fixated which is at the relevant AOI(s).

3. *AOI fixation sequence* describes the sequence of AOI fixations during one trial. We limited this sequence to seven fixations, starting with trial onset (not counting fixations to the fixation mark). We encoded a fixation sequence of f fixations over q AOIs as a $q \times f$ binary matrix, where each column of the matrix had a 1 in the position corresponding to the AOI which was fixated, and zero otherwise. If there were fewer than f fixations, the last columns were set to 0. This binary encoding of the fixation sequence allowed us to describe any sequence of fixations without imposing an ordering of the AOIs. In addition, the fixation sequence was represented as a sequence of relevant and non-relevant AOI fixations. This representation yielded a $2 \times f$ binary matrix, in which each column had a 1 in the first row if a relevant AOI was fixated or a 1 in the second row if a non-relevant AOI was fixated. If there were less than f fixations, the last columns were set to 0. The analysis showed that the latter representation was

more informative in some cases. Note that it was necessary to use a pair of binary variables to encode each fixation of the latter representation to account for three cases: fixation at a relevant AOI or non-relevant AOI, and fewer than f fixations. The number of fixations to consider as well as the start position were determined using cross validation (CV). In cross validation, the training data are separated into k partitions, and for each partition, samples are classified using a classifier that is trained with the remaining $k-1$ partitions. The percentage of correctly classified samples over all partitions is the CV accuracy. We varied the parameters (start fixation and number of fixations) and calculated the CV accuracy when using only the fixation sequence to classify samples. Using the first few fixations gave the best results, with no improvement as later fixations were included.

4. *Duration of fixations in sequence* describes the duration of each fixation in the sequence described by variable 3. This variable was encoded by an f-dimensional vector.

5. *Total distance traveled by eye* is a scalar describing the total distance traveled by the eye gaze during a trial.

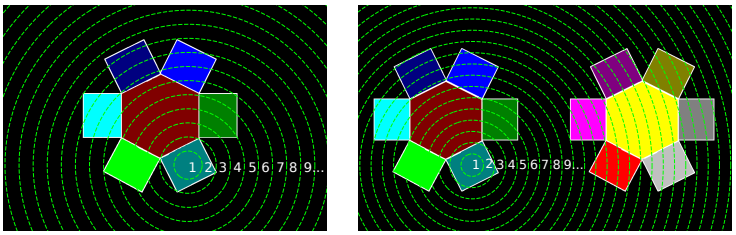
6. *Histogram of fixation distances to relevant AOI* describes how much time is spent fixated near or far from the relevant AOI(s). A histogram with h bins and an image with r relevant AOIs yielded an $h \times r$ dimensional matrix. Each column corresponds to a different relevant AOI, and each row corresponds to a particular range of distances from that AOI. The entries define the percentage of time fixated at the distance ranges, so each column sums to 1. If no fixations

occured, all values were set to 0. The number of bins was determined using CV. The bins corresponding to AOI 4 are illustrated in Fig. 4.4.

7. *Number of unique AOIs visited* is a scalar describing the total number of unique AOIs fixated during a trial. AOI revisits were not counted as new.

8. *Saccade sequence* is similar to variable 3 but describes the sequence of AOI saccades during one trial. All saccades whose targets were not to AOIs were discarded in this and all of the variables defined. The sequence was limited to seven saccades, starting at the first saccade. The number of saccades to consider as well as the start saccade were determined using CV. We encoded a saccade sequence of s saccades over q AOIs as a $q \times s$ binary matrix. Each column of the matrix had a 1 in the position corresponding to the AOI which was the target of the saccade, and zero otherwise. If there were fewer than s saccades, the last column(s) were set to 0. In addition, the saccade sequence was represented as a sequence of saccades to relevant and non-relevant AOIs. This representation yielded a $2 \times s$ binary matrix, with each column containing a 1 in the first row if saccading to a relevant AOI or a 1 in the second row if saccading to a non-relevant AOI. If there were fewer than s saccades, the last column(s) were set to 0.

9. *Relative number of saccades to an AOI* is the saccade analogue of variable 1 and describes the relative number of saccades to the AOIs during one eye movement. An image with q AOIs yielded a q-dimensional feature vector with each entry counting the number of saccade targets at the corresponding AOI.



(a) Single object (b) Object pair

Figure 4.4: Illustration of variable 6, the distance histogram bin (DHB). AOI 4 DHB regions have been numbered for clarity. DHB 1 to 35 describe the percent of time fixating within the corresponding bins. For example, DHB 1 = 0.5 means half the total fixation time was within the first bin. Bin sizes were determined using cross validation.

The vector was normalized by the sum of all entries such that the entries added to 1 unless there were no saccades. In that case, all entries were set to 0.

10. *Fixation latency to relevant AOI* describes the delay before fixating at a relevant AOI. It was defined as the duration from the trial start to the beginning of the first fixation at a relevant AOI.

11. *Saccade latency to relevant AOI* describes the delay in seconds before a saccade to a relevant AOI. It was defined as the duration from the trial start to the end of the first saccade to a relevant AOI.

Thus, eye movements were represented by a *feature vector* $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$ whose d entries correspond to the variables described. Each feature x_i was normalized to zero mean and unit variance over the entire dataset. In addition, each $\mathbf{x}$ was associated with a class label, $y \in \{0, 1\}$. For clarity, *features* denote the entries of the feature vector which encodes the eye tracking variables, while *variables* correspond

to the measures of the eye tracking enumerated above. Therefore, d is much larger than 11, because encoding certain variables requires multiple feature values. Note that d was the same for all feature vectors corresponding to images having the same number of AOIs and relevant AOIs because a fixed number of fixations and saccades were analyzed.

In the case of a single category object having one relevant AOI, variable 2 is identical to one of the values of variable 1. Therefore, after extracting all variables from the gaze data of all participants, we did a simple redundancy check to eliminate cases of identical valued features. For features x_i, x_j to be identical, they must mirror each other over *all* feature vectors for a particular category condition and within either the learning or testing phase. In addition, the information encoded by several of these features overlaps. This over-complete representation allows us to find the encoding that is best suited to describe the categorization task. To this end, we performed variable selection on this over-complete set.

4.1.9 Variable Selection

Our goal was to identify the subset of variables from the set defined above that can best separate learners from non-learners. This was achieved using ANOVA feature selection by ranking, Naive Bayes Ranking (NBR), and L1 logistic regression (L1-LR).

ANOVA feature selection relies on a standard hypothesis test on each feature of $\mathbf{x}$. Specifically, let x_i denote the i^{th} feature of $\mathbf{x}$. Using a dataset of eye tracking feature vectors and the associated class labels, we performed a two tailed t-test of the null

hypothesis, which states that samples of x_i coming from classes 1 and 0 are independent random samples from normal distributions with equal means, μ_{i1} and μ_{i0}, respectively. The alternative says that the class means are different. We calculated the test statistics and the corresponding p-value. A low p-value means the null hypothesis is rejected with confidence. Since the goal was to find the variables which best separate the classes, the feature with lowest p-value was ranked as best. The p-values were calculated for all features $x_i, i = 1 \ldots d$, and they were ranked from best to worst according to increasing p-values.

Naive Bayes Ranking (NBR) assumes that if the labeled feature vectors can be accurately classified given a single feature, x_i, then that feature separates the two classes well. In essence, the classification accuracy is a surrogate for the class separability achieved by the particular feature. Therefore, the features are ranked from best to worst according to decreasing classification accuracy.

The Bayes classifier assigns a sample, $\mathbf{x}$, to the class having the highest posterior distribution. More formally, assume that the class-conditional density functions of a feature, given its class, $p(x_i|y)$, are modeled as normally distributed with mean and variance, μ_{iy} and σ_{iy}, respectively. Then by applying the Bayes formula, the posterior probability of class y is $P(y|x_i) = \frac{p(x_i|y)P(y)}{p(x_i)}$, where $P(y)$ is the prior of class y, and $p(x_i)$ is a scale factor which ensures that the probabilities sum to 1. In this work, we have $P(y = 1) = P(y = 0) = 0.5$, corresponding to the assumption that *a priori* a sample is equally likely to come from a learner as from a non-learner. The scale factor is the same for both classes, so it can be omitted in the classification rule. Finally,

the predicted class label, $\hat{y}$ is given by:

$$\hat{y} = \arg \max_{j \in \{0,1\}} p(x_i|y = j)P(y = j). \tag{4.1}$$

L1 *Logistic Regression* (L1-LR) is a linear classifier model, which returns a probability that a sample belongs to a particular class. It accomplishes this by modeling the natural logarithm of the ratio, or odds, of two probabilities as a linear function of $\mathbf{x}$. More formally,

$$ln\left(\frac{p(y = 1|\mathbf{x})}{1 - p(y = 1|\mathbf{x})}\right) = \mathbf{w}^T\mathbf{x} - b, \tag{4.2}$$

where ln denotes the natural logarithm. The two class probabilities are then given by

$$\begin{aligned} p(y = 1|\mathbf{x}) =& \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x} + b)}, \\ p(y = 0|\mathbf{x}) =& \frac{\exp(-\mathbf{w}^T\mathbf{x} + b)}{1 + \exp(-\mathbf{w}^T\mathbf{x} + b)}. \end{aligned}$$

The parameters, $\mathbf{w}$ and b, are estimated via Maximum Likelihood (ML) estimation. A regularization term λ is introduced to penalize large elements in $\mathbf{w}$. Using an L1-norm regularizer yields a sparse model. More formally, the regularized ML objective is,

$$\hat{\mathbf{w}} = \arg \max_{\mathbf{w},b} \sum_{i=1}^{N} logP(y_i|\mathbf{x}_i) - \lambda\|\mathbf{w}\|_1, \tag{4.3}$$

where $(\mathbf{x}_i, y_i), i = 1, \ldots, N$ are the full feature vectors and their associated labels, λ is a user determined real valued positive regularization parameter, and $\|\cdot\|_1$ denotes the L1-norm. Increasing the value of λ will result in more elements of $\mathbf{w}$ being shrunk to zero, i.e., a sparse weight vector $\mathbf{w}$. Variable selection is performed by increasing

the value of λ until a desired number of $\mathbf{w}$ elements are non-zero. The elements of $\mathbf{x}$ corresponding to the non-zero elements of $\mathbf{w}$ are the top ranked variables. These top ranked variables can then be sorted from best to worst by sorting the corresponding entries of $\mathbf{w}$ in order of descending absolute magnitude. We use the L1-LR implementation of [102].

Each method results in a ranking of the features, x_i, from best to worst. If we vectorize the indices of the t top ranked features as $\mathbf{k} = (k_1, k_2, \ldots, k_t)^T$, then after feature selection $\mathbf{x} = (x_{k_1}, x_{k_2} \ldots, x_{k_t})^T$.

4.1.10 Linear Classification

Once the important variables were identified, we used them to classify the gaze data as having originated from a learner or non-learner. This required that we train a classifier to distinguish between two classes of data. Recall that each eye movement resulted in a feature vector, or *sample* $\mathbf{x}$. A classifier defines a decision rule for predicting whether a sample is from class 0 or 1. A linear classifier was used because of its ease of interpretation [75] – the absolute model weights give the relative importance of the eye tracking variables. We illustrate in Fig. 4.5 with a 2-dimensional linear classifier model specified by $\mathbf{w}$ and b. $\mathbf{w}$ is the normal vector of the hyperplane which separates the feature space into two decision regions, and b is the distance from the origin to the hyperplane (i.e., the offset).

All samples $\mathbf{x}$ above the hyperplane are assigned to class 1 while the samples below are assigned to class 0. Data samples $\mathbf{x}$ existing on the boundary satisfy $\mathbf{w}^T\mathbf{x} - b = 0$. Therefore, samples are classified according to the sign of $\mathbf{w}^T\mathbf{x} - b$. In this example $\mathbf{w} = (-.55, .83)^T$, so the second dimension, x_2, is more informative

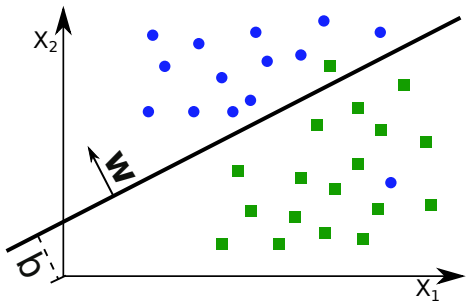


Figure 4.5: Illustration of a linear classifier. $\mathbf{w}$ is the normal vector of the hyperplane which separates the feature space into two decision regions, and b is the distance from the origin to the hyperplane. The blue circles represent samples from class 1, while the green squares represent samples from class 0. All but one of the blue circles exists on the positive side of hyperplane, and are classified correctly.

for classification. Note that in our case the feature space has not two but up to 334 dimensions, depending on the cut-off for variable selection.

Several varieties of linear classifiers exist. In this work, we used the Bayes classifier with equal covariances, L1-LR, and the Support Vector Machine (SVM) algorithm.

Bayes with equal covariances (Bayes): When both classes are assumed to be multivariate normally distributed with the same covariance Σ, means μ_1 and μ_0, and equal priors, the Bayes classifier decision boundary is a hyperplane given by $\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_0)$ and $b = \frac{1}{2}\mathbf{w}^T(\mu_1 - \mu_0)$ [29].

L1 *Logistic regression:* Recall that L1-LR yields a probability that a sample belongs to a particular class. It uses the model of Equation (4.2), where $\mathbf{w}$ defines the normal of the hyperplane, and the sign of $\mathbf{w}^T\mathbf{x} - b$ determines the class label.

Support Vector Machine: SVM is a linear classifier which maximizes the margin between two classes of data [16]. In the case that the training samples are perfectly separable by a hyperplane, we can find $\mathbf{w}$ and b such that the data satisfies the following constraints,

$$\mathbf{x}_i^T\mathbf{w} - b \geq 1 \text{ for } y_i = 1, \tag{4.4}$$

$$\mathbf{x}_i^T\mathbf{w} - b \leq -1 \text{ for } y_i = 0. \tag{4.5}$$

Essentially, these constraints specify that the samples from the different classes reside on opposite sides of the decision boundary. The margin between the classes, defined by $\frac{2}{\|\mathbf{w}\|_2}$ where $\|\cdot\|_2$ defines the L2-norm, is then maximized subject to the above constraints. The dual formulation of the constrained optimization problem results in a quadratic program for $\mathbf{w}$ and b. In the case that samples from each class are not linearly separable, a penalty is introduced to penalize the amount that a sample is on the wrong side of the hyperplane. Again, the dual formulation results in a quadratic program for $\mathbf{w}$ and b. We used the implementation of [19].

4.1.11 Classification Accuracy

The classification accuracy used for adults was the leave-one-subject-out cross-validation (LOSO-CV) accuracy. In LOSO-CV, the samples belonging to one participant are sequestered, and the remaining samples are used to train the classifier. The sequestered samples are then classified with the learned classifier, and the procedure is repeated for every participant in the database. The total number of correctly classified samples divided by the total number of samples is the LOSO-CV accuracy.

The classification accuracy used for infants was the leave-one-experiment-block-out cross-validation (LOBO-CV) accuracy. This alternative accuracy measure makes

more effective use of the eye movement data when the sample size is very small. In LOBO-CV, the samples belonging to one experiment block are sequestered, and the remaining samples are used to train the classifier. The sequestered samples are then classified with the learned classifier, and the procedure is repeated for every block in the database. The total number of correctly classified samples over the total number of samples is the LOBO-CV accuracy.

4.2 Results

4.2.1 Adult Experiment

We first labeled the adult trials as category learner or non-learner. This resulted in 728 learning class samples and $1,256$ non-learning class samples for the learning phase, and 473 learning class samples and 601 non-learning class samples for the testing phase in the category A or B category learning condition. There were 496 learning class samples and $1,568$ non-learning class samples for the learning phase, and 323 learning class samples and 717 non-learning class samples for the testing phase in the category C or D category learning condition. The indeterminate samples were not analyzed. We then extracted the eye tracking variables from each trial's gaze sequence. Each labeled data sample resulted in a 182-dimensional feature vector for the learning phase samples, and a 334-dimensional feature vector for the testing phase samples.

We applied the variable selection algorithms to identify the most important variables for separating learners from non-learners, and validated those variables using the three linear classifiers. The LOSO-CV error is reported as a function of the number of top features used for classification in Fig. 4.6. Recall that the features encode the

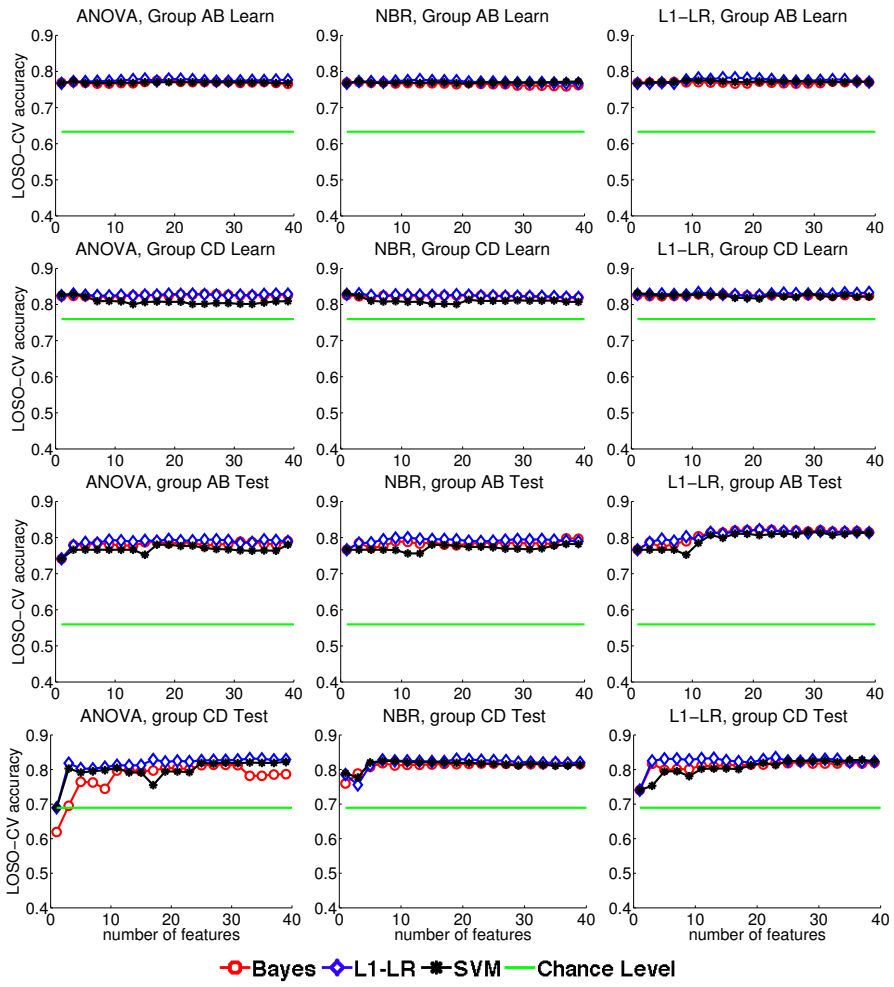


Figure 4.6: Leave one subject out cross-validation accuracy for adult subjects as a function of the number of top ranked variables used for classification. The first two rows show results for the learning phase of the experiments (categories AB and CD, respectively). The last two rows show the results for the testing phase of the experiments (categories AB and CD, respectively). ANOVA, NBR, and L1-LR correspond to ANOVA feature selection, Naive Bayes feature selection, and L1 penalized logistic regression feature selection, respectively. AB and CD correspond to category object A or B and C or D respectively. In almost all cases, the classification accuracy was near the maximum after including very few features and did not change much when including more. Chance level is plotted as the accuracy resulting from classifying each sample as the most common class.

eye tracking variables. The results show that a very small numbers of features yield a high classification rate, and including more features does not improve the accuracy. Using the top five ANOVA features and the Bayes classifier, we achieved LOSO-CV accuracies of 77% and 82% in the AB and CD learning conditions, respectively. Similarly, we achieved 78% and 76% accuracy in the testing conditions.

The stable performance beyond just a few features suggests that a small number of variables is sufficient for discriminating learners and non-learners. The top five variables for ANOVA, NBR, and L1-LR are listed in Table 4.2. We boldfaced the variables that were consistently ranked in the top five variables across both the category A or B and category C or D conditions and all feature selection algorithms. We underlined variables that were consistently ranked in the top five variables by at least two of the three features selection algorithms and across category A or B and category C or D conditions. Note that AOI 4 for the category A or B condition is equivalent to AOI 6 in the category C or D condition. The consistent top variables in the learning condition were *latency to a fixation at the relevant AOI*, *density of fixations at the relevant AOI*, and *first and second fixation at the relevant AOI*. The top variables in the testing condition were *first, second, and third fixations* as well as *second saccade*.

4.2.2 Infant Experiment

We first labeled the infant trials as category learner or non-learner. This amounted to 135 learning class samples and 137 non-learning class samples for the learning phase, and 40 learning class samples and 40 non-learning class samples for the testing phase in the category A or B category learning condition. The C or D category

	Learning condition		
	ANOVA	NBR	L1-LR
A or B	1. **Lat to fix rel AOI**	**Lat to fix rel AOI**	**Lat to fix rel AOI**
	2. **Den of fix at rel AOI**	**Den of fix at rel AOI**	**Den of fix at rel AOI**
	3. rel AOI, DHB 2	rel AOI, DHB 2	4^{th} fix at rel AOI
	4. 2^{nd} fix at rel AOI	2^{nd} fix at rel AOI	3^{rd} fix at rel AOI
	5. 1^{st} fix at rel AOI	1^{st} fix at rel AOI	5^{th} fix at rel AOI
C or D	1. **Lat to fix rel AOI**	**Den of fix at rel AOI**	**Den of fix at rel AOI**
	2. **Den of fix at rel AOI**	**Lat to fix rel AOI**	**Lat to fix rel AOI**
	3. rel AOI, DHB 5	rel AOI, DHB 2	rel AOI, DHB 2
	4. 1^{st} fix at rel AOI	1^{st} fix at rel AOI	Den of fix at AOI 1
	5. 1^{st} fix at non-rel AOI	2^{nd} fix at rel AOI	Den of fix at AOI 2

	Testing condition		
	ANOVA	NBR	L1-LR
A or B	1. 3^{rd} fix at non-rel AOI	2^{nd} fix at non-rel AOI	2^{nd} fix at non-rel AOI
	2. 2^{nd} fix at non-rel AOI	2^{nd} sac to non-rel AOI	1^{st} fix at non-rel AOI
	3. 2^{nd} sac to non-rel AOI	1^{st} fix at non-rel AOI	3^{rd} fix at non-rel AOI
	4. Lat to fix rel AOI	Duration of 3^{rd} fix	2^{nd} sac to non-rel AOI
	5. Lat to sac rel AOI	3^{rd} fix at non-rel AOI	1^{st} sac to non-rel AOI
C or D	1. 4^{th} fix at non-rel AOI	Den of fix at rel AOI	2^{nd} sac to non-rel AOI
	2. 3^{rd} fix at non-rel AOI	1^{st} fix at non-rel AOI	1^{st} fix at non-rel AOI
	3. Lat to fix rel AOI	Den of fix at AOI 13	Den of fix at rel AOI
	4. 2^{nd} sac to non-rel AOI	1^{st} fix at rel AOI	2^{nd} fix at non-rel AOI
	5. Number AOIs fixated	2^{nd} fix at non-rel AOI	3^{rd} fix at non-rel AOI

Table 4.2: Adult Experiment: The variables above were determined most relevant during the category learning and category discrimination phases of the adult experiment. The bold face entries show variables that were consistently determined most relevant using all feature selection algorithms and on two separate category object conditions. The underlined entries show variables that were determined most relevant by at least two feature selection algorithms and across both category conditions. ANOVA, NBR, and L1-LR correspond to the different feature selection algorithms. AOI 4 is relevant in the category A or B condition, and corresponds to AOI 6 in the category C or D condition. We use the following shorthand convention: fixation (fix), saccade (sac), relevant (rel), density (den), latency (lat), distance histogram bin (DHB).

learning condition resulted in 139 learning class samples and 127 non-learning class samples for the learning phase, and 40 learning class samples and 40 non-learning class samples for the testing phase. As in the adult experiment, the indeterminate samples were not used. After labeling the data and extracting the variables from each gaze sequence, each sample resulted in a 334-dimensional feature vector for both the learning and testing phase samples.

The three linear classifiers discussed above were applied to determine the LOBO-CV error as a function of the number of top features selected by the three different feature selection algorithms. The results are shown in Fig. 4.7, where we see that classifying infants requires significantly more variables than the adult case. This is to be expected because of the diffuse looking pattern typical of babies. Using the top ten ANOVA variables and the Bayes classifier, we achieved LOBO-CV accuracies of 61% and 64% in the AB and CD learning conditions, respectively. Similarly, we achieved 66% accuracy in the testing conditions. The top ten infant variables are shown in Table 4.3. The bold face entries were consistently selected by all three feature selection algorithms across both category conditions. The consistent top variables in the learning and testing conditions were *density of fixations* and *DHB*, which describes the density of fixations at different distances from the relevant AOI(s). The *fourth* and *sixth fixations* were also relevant in the testing condition.

4.2.3 Comparing Infants to Adults

The above results raise a new question. How similar are the attention models of adults and infants? Specifically, since the infant data are so noisy, can we use the adult model to improve on the infant one? To test this, we used the adult Bayes

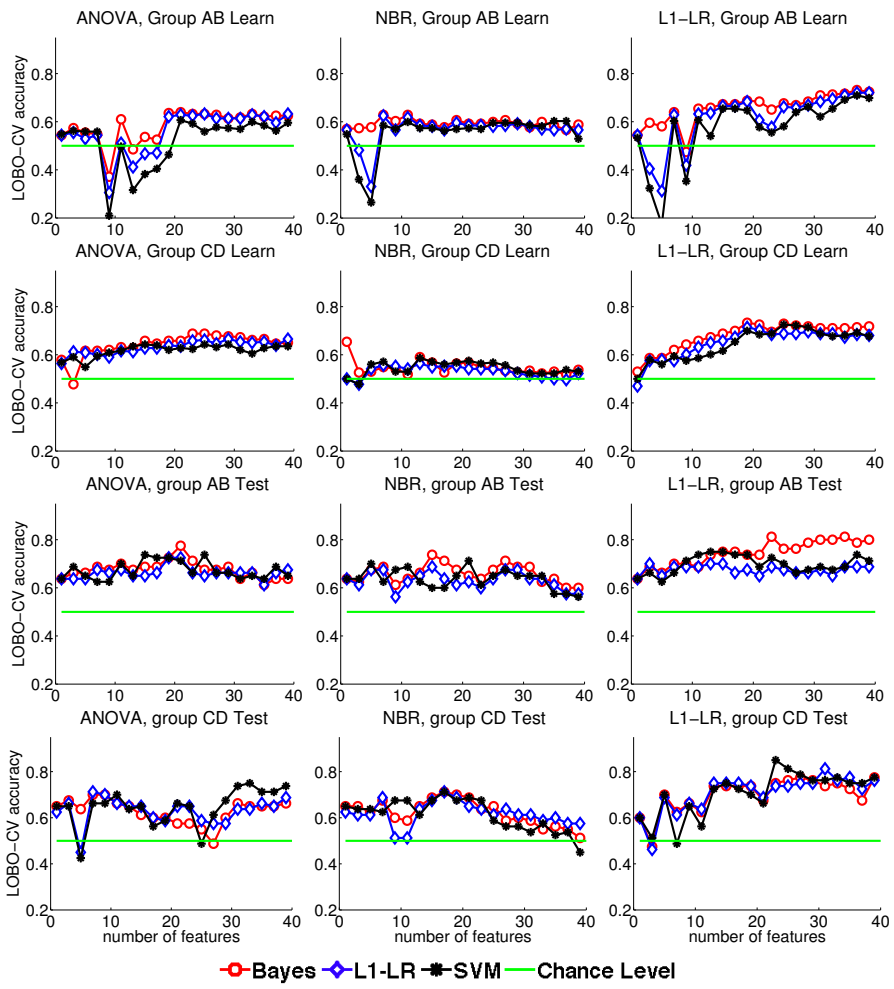


Figure 4.7: Leave one experimental block out cross-validation accuracy for infant subjects as a function of the number of top ranked variables used for classification. We use the same conventions of Fig. 4.6.

Learning condition

	ANOVA	NBR	L1-LR
A or B	1. **Den of fix** at AOI 10	**Den of fix** at AOI 2	**Den of fix** at AOI 10
	2. 3^{rd} fix at AOI 10	**Den of fix** at AOI 10	1^{st} sac to AOI 5
	3. AOI 11, **DHB** 5	AOI 11, **DHB** 5	**Den of fix** at AOI 1
	4. Den of sac to AOI 10	AOI 11, **DHB** 35	2^{nd} fix at AOI 1
	5. AOI 4, **DHB** 20	AOI 11, **DHB** 22	**Den of fix** at AOI 2
	6. 2^{nd} fix at AOI 10	Den of sac to AOI 2	AOI 11, **DHB** 5
	7. 2^{nd} fix at AOI 1	**Den of fix** at AOI 1	AOI 11, **DHB** 22
	8. **Den of fix** at AOI 1	**Den of fix** at AOI 9	AOI 11, **DHB** 16
	9. **Den of fix** at AOI 2	AOI 4, **DHB** 7	2^{nd} sac to AOI 3
	10. AOI 11, **DHB** 22	AOI 4, **DHB** 12	**Den of fix** at AOI 3
C or D	1. AOI 13, **DHB** 5	4^{th} fix at AOI 2	AOI 13, **DHB** 5
	2. AOI 6, **DHB** 21	3^{rd} sac to AOI 2	AOI 6, **DHB** 21
	3. **Den of fix** at AOI 13	1^{st} fix at AOI 5	3^{rd} sac to AOI 10
	4. AOI 13, **DHB** 2	3^{rd} sac to non-rel AOI	**Den of fix** at AOI 13
	5. 3^{rd} sac to non-rel AOI	3^{rd} fix at AOI 6	4^{th} fix at AOI 10
	6. 1^{st} fix at AOI 14	3^{rd} fix at AOI 9	3^{rd} sac to non-rel AOI
	7. 1^{st} fix at non-rel AOI	AOI 6, **DHB** 21	4^{th} fix at AOI 5
	8. **Den of fix** at AOI 14	AOI 13, **DHB** 5	1^{st} fix at AOI 14
	9. AOI 6, **DHB** 8	1^{st} fix at non-rel AOI	4^{th} sac to AOI 1
	10. 4^{th} fix at AOI 5	4^{th} fix at AOI 4	AOI 13, **DHB** 2

Testing condition

	ANOVA	NBR	L1-LR
A or B	1. **6^{th} fix** at non-rel AOI	**6^{th} fix** at non-rel AOI	**6^{th} fix** at non-rel AOI
	2. AOI 4, **DHB** 20	AOI 4, **DHB** 20	AOI 4, **DHB** 8
	3. AOI 4, **DHB** 8	**Den of fix** at AOI 10	**Den of fix** at AOI 13
	4. **4^{th} fix** at AOI 7	Den of sac to AOI 10	**4^{th} fix** at AOI 7
	5. Den of sac to AOI 10	**4^{th} fix** at AOI 7	AOI 4, **DHB** 20
	6. AOI 4, **DHB** 10	AOI 4, **DHB** 8	AOI 11, **DHB** 16
	7. **Den of fix** at AOI 10	AOI 4, **DHB** 10	7^{th} fix at AOI 7
	8. 7^{th} fix at AOI 10	number AOIs fixated	AOI 4, **DHB** 10
	9. 1^{st} sac to AOI 10	1^{st} fix at AOI 13	Den of sac to AOI 10
	10. 1^{st} fix at AOI 1	2^{nd} fix at AOI 14	**Den of fix** at AOI 1
C or D	1. **Den of fix** at AOI 7	**Den of fix** at AOI 7	3^{rd} sac to AOI 3
	2. 3^{rd} sac to AOI 3	3^{rd} fix at AOI 7	**Den of fix** at AOI 7
	3. 1^{st} fix at AOI 7	**6^{th} fix** at AOI 7	**4^{th} fix** at AOI 10
	4. **4^{th} fix** at AOI 10	Duration of 2^{nd} fix	**4^{th} fix** at AOI 12
	5. 3^{rd} fix at AOI 7	AOI 13, **DHB** 2	2^{nd} fix at AOI 1
	6. **6^{th} fix** at AOI 7	3^{rd} sac to AOI 3	2^{nd} sac to AOI 8
	7. 2^{nd} fix at AOI 1	1^{st} fix at AOI 7	2^{nd} fix at AOI 2
	8. 2^{nd} fix at AOI 2	**4^{th} fix** at AOI 7	**6^{th} fix** at AOI 8
	9. 1^{st} sac to AOI 10	**4^{th} fix** at AOI 10	2^{nd} sac to AOI 1
	10. AOI 13, **DHB** 12	AOI 6, **DHB** 7	**Den of fix** at AOI 1

Table 4.3: Infant Experiment: The variables above were determined most relevant during the category learning and category discrimination phases of the infant experiment. The bold face variables were consistently selected. We use the same conventions as Table 4.2. DHB variables correspond to density of fixating at different distances from the AOI. Larger DHBs correspond to bins that are further from the AOI.

classifier model trained with the top five variables from ANOVA to predict if infants were learners or non-learners. This was done only for the testing phase, because the testing phase images for adults and infants are similar so that the extracted variables correspond. Infants were classified with 49% accuracy in the category A or B condition. Infants were classified with 51% accuracy in the category C or D condition. These chance performance of the adult model identifying infant learners suggests that adults and infants attend to category objects differently. The remaining challenge is to examine the generality of this finding by testing a broader set of categories.

4.3 Discussion

The analysis demonstrates that the proposed methodology for identifying relevant eye tracking variables is viable. We can predict if adults have learned a category based on a very small number of top ranked eye track variables. Furthermore, there is strong agreement between the different ranking approaches about which variables are most important. Specifically, the consistently top ranked variables in the learning condition were *latency to a fixation at the relevant AOI*, *density of fixations at the relevant AOI*, and *first and second fixations at the relevant AOI*. The consistently top ranked variables in the testing condition were *first, second, and third fixations*, and *second saccade*. These results suggest that during learning, adult category learners focus their attention on the relevant category features. The results also suggest that adult category learners make discrimination judgments within the first few fixations.

The infant data analysis also demonstrated that we can predict category learning, but it requires a larger number of variables. Again, there was agreement between the

different ranking approaches about which variables are most important. The consistent top variables in the learning and testing conditions describe the *fixation density* at different AOIs and at different distances from the relevant AOIs. The *fourth* and *sixth fixations* were also relevant in the testing condition. These results suggest that for infants, the pattern of fixations over the entire object is more informative than the amount of time fixating the relevant AOI. Therefore, it appears that whereas category learning in adults is marked by focused attention to category-relevant features, category learning in infants is marked by more diffuse attention coupled with exploration of multiple areas of interest.

Finally, we showed that the adult model does not predict infant category learning. We address these findings in the next section.

4.3.1 Why were the best variables different for infants and adults?

There is an important difference between the variable selection results of the adult experiment versus the infant experiment. Namely, while adult learners are identified readily with a small set of variables emphasizing early looks at the relevant AOI(s), infant learners are better identified based on their pattern of fixating over the trial. We propose an explanation based on the goals of adult versus infant participants.

Although the experiment stimuli were the same for adults and infants, there were fundamental differences in the design of the experiments. Namely, the objectives during the experiment were different for adults versus infants. In the case of adults, the participants were given a particular task: learn how to identify a member of a given category from a set of exemplars, then identify a member of this category from a pair of objects. Therefore, the adults' goal was to learn the category object as quickly

as possible given the limited number of training examples, such that discrimination could be performed accurately during the testing phase. Given this goal, it was reasonable that the consistently selected variables were associated with relevant AOI fixation density as well as early looks (see Table 4.2).

In the case of infants, we used sound and motion to draw the infants' attention to the relevant AOI in hopes that he or she learned to identify the category object. Then, we assumed that if the category was learned, the infants would show a preference for either the learned category or novel category during the discrimination phase. To this uncertainty, we ought to add the large amounts of random movements of the infant's gaze. As we see in our results, a larger set of variables is required to reliably distinguishing learners from non-learners. In addition, while fixation density is important, the emphasis is not on fixating the relevant AOI.

4.4 Conclusion

We have developed a methodology for automatically identifying eye tracking variables relevant to a given task. Previous research has relied on ad-hoc techniques to determine which variables should be analyzed in a particular study. Instead, we used statistical methods to find the important variables in an over-complete set of variables. The efficacy of the approach was verified with an adult and infant categorization study. The variables determined most relevant for adults emphasize looking at the relevant AOI(s) longer, and earlier during the categorization tasks. This result is satisfying for two reasons: 1) It is expected that category learners quickly focus their efforts on the relevant AOI(s), and 2) these variables coincide with the variables

proportion fixation time and *relative priority* of previous eye tracking category learning studies such as [95]. The variables determined most relevant for infants emphasize the overall pattern of fixating the object. This result is also satisfying because infants are expected to explore objects.

Note that the important variables were verified by the *task* and *stimuli* described. Altering these parameters may result in different important variables. By comparing the important variables among different tasks and stimuli, we can further dissociate which eye tracking variables are linked to specific processes during categorization. Our method provides a technique for identifying the most relevant variables in each of these cases.

CHAPTER 5

VARIABLE SELECTION IN THE PERCEPTION OF FACIAL EXPRESSIONS OF EMOTION

This chapter is concerned with characterizing the way people recognize prototypical facial expressions of emotion. By prototypical emotions, we refer to the six basic emotions hypothesized to be recognized across cultures [32]: happiness, sadness, anger, disgust, surprise, and fear, plus neutral meaning no emotion is expressed. The previous works in modeling the face appearance and exploring categorization through eye tracking are the foundation for modeling emotion recognition that is developed here. We apply ideas from those works to understand, through eye tracking, the important features used in the perception of emotion from expressive faces.

Researchers have been concerned with the expression of emotion in faces and their subsequent recognition for over 150 years. In his pioneering work, *Mécanisme de la Physionomie Humaine* [25], Duchenne gives a meticulous account of the mechanism of facial expression production in humans. Duchenne used electrodes to stimulate individual face muscles to determine which muscles produced different facial expressions of emotion. By moving a single feature such as raising the eyebrows, he noticed the entire face appears to change expression. He also noted that certain muscles could not be voluntarily controlled, but moved involuntarily in response to an emotional

state. He attributed this to an emotion of the *soul*. Duchenne argued that the ability to express specific emotions by contracting the same sets of muscles is innate and universal, in a manner defined by the Creator.

Just a decade later in 1972, Darwin published his influential work, *The Expression of the Emotions in Man and Animals* [24], where he sought to characterize the movements of features of the body that characterize certain states of mind. He argued an evolutionary perspective, that the facial expressions of emotion are rooted in a common ancestor. Through studies of small children, the blind, the insane, and different cultures, he argues why many emotional expressions of the body are inherited habitual actions that were consistently associated with some state of mind, or emotional state.

These habitual actions are similar to reflexes. They may have served some purpose at one point, and are inherited. For example, dogs will circle around on the carpet before going to bed as if to make a recess to sleep in. This action is probably inherited from previous generations of dogs that lived outdoors. There are also some *opposite actions*, actions of the body that directly oppose some habitual action, and become themselves habitual and inherited. Finally, actions due to *excessive nerve force*, such as shaking during terror, can also be inherited.

Ekman built upon the ideas of Duchenne and Darwin in his own research of emotional facial expression. He argued through cross-cultural studies, among others, that some emotional expressions are innate, and recognizable across cultures: happiness, sadness, anger, disgust, surprise, and fear [32]. One criticism of cross cultural studies in testing the innateness of the facial expressions of emotion is that many tested

cultures have been exposed to western media, allowing them to learn the expressions. Therefore, Ekman conducted studies with members of a pre-literate culture in New Guinea that had not been exposed to Western media. He read participants a story having some emotional content, then showed them three photographs of models expressing different prototypical facial expressions. Participants identified the faces displaying the emotional relevant to the story, although they had difficulty distinguishing fear from surprise.

A significant amount of developmental research has also been conducted to determine if the perception of emotional expressions is innate, or learned. Several studies have suggested that infants discriminate between different facial expressions of emotion as early as 36 hours after birth [40]. In this study 74 neonates averaging 36 hours were habituated to the facial expressions of happy, sad, and surprise by a model who fixed each expression on her face as the baby looked at her. An observer unaware of the model's expression was able to guess which expression was shown to the neonate based on the neonate's imitation of the model's facial expression. Visual fixation significantly reduced later in the trial, and significantly increased once a new expression was shown, providing evidence for discrimination. Another study by Young-Browne *et al.* [130] found evidence for discriminating sad, happy, and surprise by 3-month-olds using an infant controlled habituation-recovery procedure.

There are important concerns regarding the findings of those studies. Some authors have noted that it is unlikely that infant visual system is developed enough to perceive the expression relevant face information much younger than 4 months [17, 83]. Second, we cannot be certain that the infant dehabituates due to a novel emotion *category*, or to some noticed change in the face, such as a mouth opening or

closing. To test for the recognition of emotional expression categories, authors have familiarized infants to facial expressions posed by one or more models, and tested the discrimination of emotions posed by yet another model. Using this approach, evidence has been found for the categorical discrimination of emotional expressions from as early as three months. Nelson and Dolgin found evidence for discriminating happy versus fear in 7-month-olds [85], although discrimination was only evident when the infants first habituated the happy face. Such asymmetry was found for other cases of discriminating anger. For example, Schwartz *et al.* [104] found that 5-month-olds could discriminate anger, fear and sadness, except when anger was the novel stimuli. Unlike other studies, they failed to show evidence for discriminating between joy, anger, and interest in that age group. Caron *et al.* [17] tested $4\frac{1}{2}$-, 6-, and 7-month-olds ability to discriminate happy from surprise across identities. The youngest group did not generalize discrimination across identities, while the 6-month group discriminated only when habituated to happy. The oldest group generalized discrimination across identity. In summary, the developmental literature suggests that at least some aspects of emotion recognition are innate, with experience playing an important role over the first two years [84].

Besides the question of nature versus nurture, another fundamental question remains. Specifically, there is debate about whether emotions are actually perceived is distinct categories, or vary continuously across underlying dimensions. The first school of thought gives rise to the categorical model which assumes that each emotion is associated with a dedicated neural system that is active in response to the emotion or when expressing it. Researchers have proposed a set of emotion categories from as early as the 1920s and tested participants' judgment of a variety of emotion face

stimuli using a set of possible words [1, 41, 125, 90, 120, 87]. It is not possible to conclude that emotions are categorical, but by summarizing the findings from these early studies we can conclude that seven posed emotion categories can be recognized from photographs: happiness, surprise, fear, anger, sadness, disgust, and interest [31]. A short review of the categorical model along with supporting evidence can be found in [129].

Soon after the theory of emotion categories, emerged the emotion dimension theories [103, 87, 57, 42, 91]. This alternative approach, called the continuous model, assumes that emotions arise from continuous overlapping neural systems which define common dimensions such as pleasantness and intensity along which all emotions vary. These works taken together suggest that there may be two or three dimensions underlying the emotions, but there is disagreement about what the dimensions should be. A short review of the continuous model along with supporting evidence can be found in [91].

We do not address the question of whether a categorical or continuous model better explains human behavior. Instead, we address the more specific questions of where emotion information is evident in the face, and how the features used for recognition of emotional expression categories change over development. Previously, research has shown that adults exhibit different looking patterns when looking at different emotional facial expressions [30]. An example is shown in Fig. 5.1 for the 7 prototypical emotions. These maps are obtained by standardizing the weighted fixation density maps to zero mean and unit variance, averaging across all trials for each emotion separately, then subtracting the overall mean of all standardized fixation density maps. The weighted fixation density maps describe the fixation pattern over

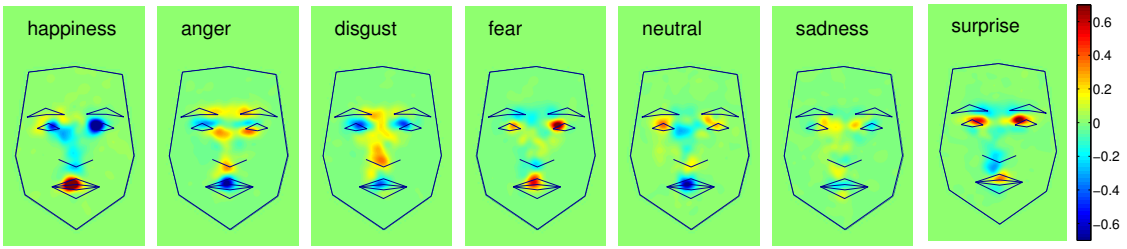


Figure 5.1: Marginal fixation density maps for the different emotions for adults with respect to the overall mean.

trials, such that areas fixated for longer have a large value. Gaussian smoothing is applied to account for slight differences in fixation positions.

Since gaze is linked to attention allocation over time, we hypothesize that the differences in gaze sequence may reflect different relevant face features when recognizing different emotions. If the gaze sequence differs systematically for different emotions, then the gaze sequence should be diagnostic of the stimuli's emotion category. We do an eye tracking study in order to understand what drives these differences in gaze and to answer the following key questions about emotion recognition.

1. What features of the face are relevant to emotion recognition? Are these features different for different emotion categories?

2. Which aspects of the gaze such as fixations, saccades, and sequence are most informative about the emotion stimulus.

3. Is a particular segment of the face viewing time more informative than the others?

4. How does that pattern of looking change over development?

These questions are concerned with the cognitive space of emotion recognition. The important features and aspects of gaze suggest appropriate dimensions for representing the emotions.

The analysis requires that we expand the methodology developed for the simple category objects of the previous chapter. Those category objects are well defined in that a single specific rule based on a single AOI defines the object category. However, the dimensions for emotion recognition from expressive faces are not as obvious. It is known that the face muscles contract and relax in ways that are characteristic of the different facial expressions [33], and identifying these muscular configurations is one computational approach to emotion recognition [119]. However, recent research has suggested that configural features may play a significant role in emotion perception [86], with shape and texture being the other two alternatives we must consider. Therefore, our analysis makes use of the gaze pattern over the entire face as well as the first order sequence, which may relate to relational information extraction. We also expand our standard toolbox of linear classifiers to include cascade based classifiers which will better model stage-wise recognition approaches. The results suggest a holistic face representation for recognizing emotional facial expressions, with certain areas having influence on gaze according to the emotion category.

5.1 Methodology

5.1.1 Participants

This study was approved by a Behavioral and Social Sciences IRB at The Ohio State University. Thirty-four adults participated in the study. They were introductory psychology students at The Ohio State University who participated for course

credit. They were given sufficient information and signed a consent form before freely participating in the study. Twenty-one infants participated in the study. Their parents were given sufficient information before providing written consent. All parents reported their infants to be developing typically and without known health problems.

5.1.2 Materials

Adult eye movement was recorded by an Eyelink 1000 Tower (SR Research, Ontario, Canada) at 1000Hz. Participants sat with with their chin resting on a mount and their eyes approximately 60cm from a 12×16 inch display, spanning approximately 28° vertical by 37° horizontal visual angle. A simple eye dominance test was performed before the experiment, and only the dominant eye was tracked.

Infant eye movement was recorded by an Eyelink 1000 Remote Arm (SR Research, Ontario, Canada) at 500Hz. Infant participants were buckled into a car seat with their eyes approximately 60cm from a 10.5×13.25 inch display, spanning approximately 25° vertical by 31° horizontal visual angle. Typically only the right eye was tracked, unless there was difficulty during calibration. In such a case, the left eye was tracked if calibration was successful with the left eye.

5.1.3 Experiment 1 - Adults

In the adult experiment, participants were first instructed that they would be presented with a series of emotional face photographs, and asked to identify which emotion was expressed after each photograph. Next, we showed participants a screen with a labeled example of all 7 emotions (happiness, sadness, anger, disgust, surprise, fear, and neutral). Subjects reviewed the examples and pressed a button on a hand held controller when they were ready to proceed.

Before each block, the camera was calibrated on a 9 points grid and mean error was verified to be less than 1° of visual angle. Every stimulus image was preceded by a drift correction screen, which allowed to eye tracker to compensate for small errors and let the experimenter re-calibrate the camera if the accuracy reduced. The drift correct screen was gray with a centered cross-hair. Subjects were instructed to fixate the cross-hair. An experimenter sitting at the eye tracking computer monitored the gaze position and pressed enter when the subject was fixating the cross-hair to prompt the stimulus display.

Adult stimuli were centered and contrast normalized grayscale face photographs of expressive faces posed by models of different gender and races, spanning 23° vertical by 15° horizontal visual angle. The images were displayed for 2000 ms, immediately followed by the response screen. The top of the response screen always displayed the question "What best describes the emotion just seen?", with the following options arranged in a circle: happiness, sadness, disgust, neutral, fear, surprise, and anger. The position of the seven choices was constant within subjects, but randomized across subjects. Subjects responded by looking at their answer to highlight it then selected the highlighted answer by pressing any button on a hand held controller. This allowed subjects to respond without looking away from the screen or moving their head.

The experiment began with a practice block where the participants saw an example of all seven emotions and become familiar with the method of answer selection. Then, there were 182 trials consisting of 26 different examples from all 7 emotion categories. The order of their presentation was randomized across subjects. The experiment was split into 4 blocks of approximately 46 trials each. Before each block, participants were given a short break to stretch and and rest their eyes.

5.1.4 Experiment 2 - Infants

The infant experiment used an infant controlled paired discrimination design to test for discrimination of the emotion categories across infants of different ages. The neutral category was removed to account for the limited duration of infant cooperation while testing a broad set of emotion categories. Photographs were presented in pairs so that that familiarization and discrimination trials had uniform layout. The camera was calibrated at the beginning of the experiment on a 3 points grid where a looming audible circle attracted participants' attention to the calibration points.

The experiment proceeded through six blocks, where each block checked for participants' ability to discriminate all the emotional expression categories from one familiarized category exemplar. The order of the emotions to familiarize in each block was randomized across subjects.

Each block began with drift correction and an attention grabbing movie. Familiarization began when the infant fixated the center of the screen. On each familiarization trial, we showed an identical pair of images from a particular category for 4000 ms, immediately followed by a blank screen for 200 ms. The images were grayscale and illumination normalized emotional face photographs of different models of different genders expressing the six prototypical emotions. Only photographs of Caucasian models were used, and earrings were digitally removed. Each photograph spanned approximately 19° vertical by 13° horizontal visual angle. Each image was positioned at approximately 8° horizontal visual angles from the center of the screen.

To assess discrimination of emotion expression category across identity, we showed three exemplars of different identities, and repeated the sequence of three images until total looking time within the three face pairs reduced by 30% of the initial

looking time, looking dropped below 2000 ms, or five iterations of all three images were presented. We then performed drift correction to allow the experimenter to recalibrate if necessary before proceeding with the discrimination trials.

Each discrimination trial began with an attention grabbing movie drawing the infant's attention to the center of the screen. Once fixated at the center, we displayed two different images of expressive faces posed by the same model for 4000 ms. One was an exemplar from the familiarized emotional expression category while the other was a novel expression. For every novel category, we showed two separate identities modeling both the familiarized and novel category. The horizontal position of novel exemplar was counterbalanced. Familiarization was repeated after three categories (six trials), before testing the last two categories of the block (four trials).

5.1.5 Aligning Gaze

Gaze sequence is typically described as a weighted fixation density map (WFDM) which describes the overall pattern of fixating the 2D image, weighted by the amount of time spent fixating the different areas. However, there is much more information such as the fixation or saccade latency contained in the series of saccades and fixations over each trial. Therefore, similar to our previous work on identifying the variables relevant to categorization, we encode each additional variable using a gaze map (GM). Examples are shown in Fig 5.2 for the six different GMs analyzed for a single experiment trial.

The GM representation is convenient for comparing the gaze pattern across subjects and stimuli when using aligned stimuli images, such as schematic faces. One important problem is spatially aligning the gaze information when comparing the

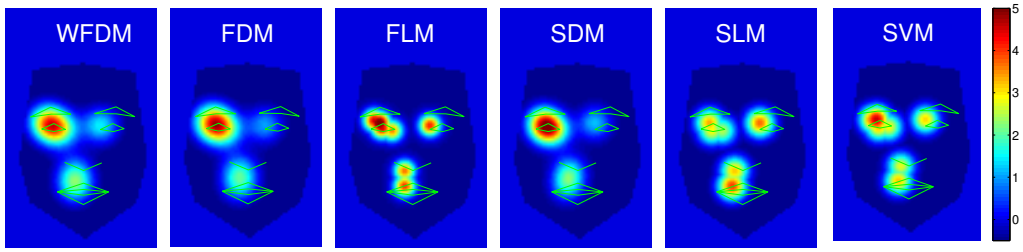
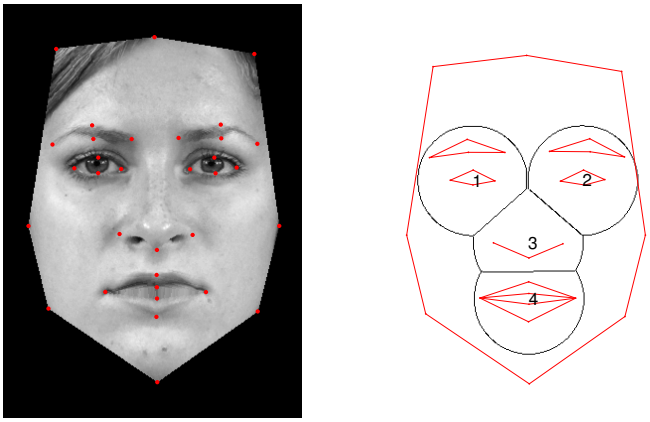


Figure 5.2: The six GMs are displayed above for a single trial. From left to right, the maps correspond to the weighted fixation density map, unweighted fixation density map, fixation latency map, saccade density map, saccade latency map, and saccade velocity map.

gaze sequences across realistic emotional face photographs. Alignment is necessary because the AOIs are not aligned in realistic face photographs so that we cannot directly compare GMs. To overcome this, researchers typically define discrete AOIs based on important anatomic landmarks and record fixations and saccades with respect to the these regions [58, 30, 88, 97]. This approach dramatically reduces the feature space dimensionality while solving the spatial alignment problem, and has the additional benefit of specifying a discrete set of states for subsequent modeling of the first order gaze sequence. The drawback is that this approach assumes a set of pre-defined AOIs. In the case of faces, the AOIs are usually restricted to the eyes, nose and mouth [58, 88, 97], or eyes and mouth [96] depending on the goal of the study.

One alternative is to warp all the face images and the corresponding gaze coordinates to a canonical face shape, then compare the gaze pattern over the aligned images. The advantage of this approach is that it is no longer necessary to pre-define AOIs because GMs can be compared directly after alignment. Furthermore, it allows the analysis to determine important face regions which may not be obvious a priori.



(a) Canonical face (b) Areas of interest

Figure 5.3: Fig. (a) illustrates an example neutral face warped to the canonical shape. Key points are shown in red. Fig. (b) shows the discrete AOIs. From one to four, the AOIs correspond to the right eye, left eye, nose, and mouth.

We apply this approach and align the faces using piecewise linear warping with 33 landmarks. An example is shown in Fig. 5.3(a).

5.1.6 Gaze variables

The GMs describe the following variables: wighted fixation density, unweighted fixation density, fixation latency, saccade density, saccade latency, and saccade velocity. Every GM is formed by first creating an image of the face region where every pixel is zero. In the weighted fixation density map, the locations of the fixations are set to the total fixation duration at those positions in milliseconds. The image is convolved with a Gaussian, and maps are normalized to zero mean and unit variance. In the unweighted fixation density maps, the locations of the fixations are incremented by one before convolution and normalization. This map essentially describes the number

of fixations in an area without accounting for the total time fixating that area. Fixation latency maps set the locations of the fixations to be the time since the previous fixation (or trial start) in milliseconds. These maps are also convolved and normalized as before. Saccade density maps are similar to the unweighted fixation density maps, except the saccade target positions are incremented by one. Saccade latency maps are similar to fixation latency maps, except that the saccade target positions are set to the time elapsed from the start of the previous saccade (or trial start) to the start of the current saccade. In the saccade velocity maps, the saccade target positions are set to the average velocity of that saccade before convolution and normalization.

5.1.7 Emotion classifiers

Linear Discriminant Analysis

As mentioned before, linear discriminant analysis (LDA) for the C class classification problem seeks to find the $C-1$ dimensional subspace to project the data which minimizes the Bayes error. In the two class problem with Normal distributions having equal covariance, the data is projected to the one dimensional space which achieves this minimum error. In our case, we wish to classify the 7 basic emotions (6+ neutral) so the data would be projected to the 6 dimensional subspace for classification. However, we have an additional goal of wanting to determine the most discriminative face areas so we would also like to project the gaze data to the optimal one dimensional subspace. LDA does not guarantee a 1 dimensional subspace which is best, so we use the algorithm proposed by Hamsici and Martinez [53] which solves this problem. The algorithm works by splitting the problem into a series of solvable convex problems. Essentially, there are a set of convex regions over which the one dimensional projection of the class means have a fixed ordering. The Bayes error over

these convex regions is also convex and given by [43] as

$$2C^{-1}\sum_{i=1}^{C-1}\Phi(\frac{\hat{\mu}_i-\hat{\mu}_{i+1}}{2}), \tag{5.1}$$

where Φ is the standard Normal cumulative distribution function and

$$\hat{\mu}_i=\frac{\mathbf{v}^T\mu_i}{\sqrt{\mathbf{v}^T\Sigma\mathbf{v}}}, \tag{5.2}$$

are the whitened ordered mean vectors projected onto the unit vector $\mathbf{v}$. The optimal $\mathbf{v}$ is found by minimizing the Bayes error for all orderings of the whitened class means, and selecting the solution with smallest error. To find the $C-1$ dimensional subspace to separate the classes, the algorithm iteratively finds the best one dimensional subspace in the nullspace of the previous solution. The samples are classified by the nearest mean in the projected space. Throughout the discussion, we refer to this algorithm as Bayes discriminant analysis (BDA) to distinguish from the standard LDA.

Decision Trees

The idea of decision trees is to iteratively split the feature space into smaller regions using straightforward decision rules in order to do classification. The classifier amounts to a cascade of simple classifier nodes, where at each stage of the cascade, a particular decision rule is evoked according to the result of the previous node. A test sample is proceeds through the cascade of decision nodes until a termination node, where the sample is assigned to a particular class. The decision nodes apply the linear classifier proposed by Hamsici in order to split the data such that the Bayes classification error is minimized at each iteration. Whereas some decision trees form decision nodes by splitting the data along one feature dimension, our algorithm splits

across the entire feature space at every decision node. We extend this idea to make use of the variety of GMs which represent the gaze on each trial. Specifically, our algorithm considers the Bayes error at every decision node for each GM separately, and the decision node relies on the GM which gives the best result at that node of the tree.

Normalized Scanpath Salience

The normalized scanpath salience (NSS) method used by Peters et. al. [89] considers the agreement of the participant's fixation positions with a saliency map. In our case, the salience maps are zero mean unit variance GMs corresponding to the fixation patterns of each emotion. Classification is performed by summing the value of each salience map at the fixation positions on one trial. The emotion viewed is classified as the one associated with the salience map which gives the largest value. Each saliency map is the average weighted fixation density map attributed to a particular emotion for all trials and all subjects.

Markov Model

A Markov model is a probabilistic model which assumes the Markov property [94]. In this model, a random variable (or set of random variables) transitions through a series of states, with certain transition probabilities. The Markov property means that the conditional probability distribution of future states only depends on the current state. Let us assume a system has S discrete states. The system will transition from state i to state j with probability a_{ij}. The a_{ij} are called the state transition coefficients. They have the properties that $a_{ij} \geq 0$ and $\sum_{j=1}^{S} a_{ij} = 1$, since the probability that a state transitions to all possible states must sum to unity. We define

the initial state i probability as p_i. Then, the probability of an observed sequence is given by the product of the corresponding initial probability and the state transition coefficients. In our case, the sequence of states correspond to the AOIs fixated during a trial.

5.2 Results

5.2.1 Adult Analysis

Effect of emotion category and face area on gaze

We performed a repeated measures ANOVA to determine if emotion category, face area, and emotion × area interaction were statistically significant factors in the looking times over different face areas. We considered the total fixation time within the four discrete regions in Fig. 5.3(b) as a percent of the total time fixated within the face area. We applied a square root transformed to the percent looking times preceding the repeated measures ANOVA analysis, since the transformed but not original data were normally distributed. Mauchly's test indicated that emotion category, face area, and emotion × area failed to conform to the assumption of sphericity (Mauchly's W $= .032, \chi^2(20) = 105.9, p < .001$; Mauchly's W $= .670, \chi^2(5) = 12.7, p = .027$; Mauchly's W $= .000, \chi^2(170) = 293.8, p < .001$). Therefore, results include Greenhouse-Geisser correction. We found the within subject factors of emotion category, face area, and emotion × area were all significant ($F(2.36, 77.9) = 6.17, p = .002, \eta_p^2 = .16$; $F(2.48, 81.75) = 10.3, p < .001, \eta_p^2 = .24$; $F(7.88, 260.1) = 11.7, p < .001, \eta_p^2 = .26$). We further considered face area as a factor on looking times within the emotion categories separately, to see if particular emotions or areas were causing the effect.

For the expression anger, Mauchly's test indicated that face area conformed to the assumption of sphericity (Mauchly's W $= .806, \chi^2(5) = 6.83, p = .234$). We found that face area was a significant factor on percent looking time ($F(3, 99) = 15.8, p < .001, \eta_p^2 = .323$). Post hoc comparisons on region in angry faces showed that participants spent less time looking at the mouth than any other area of the face (mouth vs. right eye: $p < .001$; mouth vs. left eye: $p < .001$; mouth vs. nose: $p = .007$).

For the expression disgust, Mauchly's test indicated that face area did not conform to the assumption of sphericity (Mauchly's W $= .648, \chi^2(5) = 13.8, p = .017$). Therefore, results include Greenhouse-Geisser correction. We found that face area was a significant factor on percent looking time ($F(2.40, 79.1) = 9.70, p < .001, \eta_p^2 = .227$). Post hoc comparisons on region in disgust faces showed that participants spent less time looking at the mouth than any other area of the face (mouth vs. right eye: $p = .001$; mouth vs. left eye: $p < .001$; mouth vs. nose: $p = .003$).

For the expression fear, Mauchly's test indicated that face area conformed to the assumption of sphericity (Mauchly's W $= .766, \chi^2(5) = 8.44, p = .134$). We found that face area was a significant factor on percent looking time ($F(3, 99) = 5.85, p = .001, \eta_p^2 = .151$). Post hoc comparisons on region in fearful faces showed that participants spent less time looking at the mouth than the eyes (mouth vs. right eye: $p = .046$; mouth vs. left eye: $p = .009$). In addition, people looked at the left eye more than the nose ($p = .044$).

For happy facial expressions, Mauchly's test indicated that face area conformed to the assumption of sphericity (Mauchly's W $= .743, \chi^2(5) = 9.43, p = .093$). We found that face area was a significant factor on percent looking time ($F(3, 99) =$

$4.36, p = .006, \eta_p^2 = .117$). Post hoc comparisons on region in happy faces showed that participants looked at the right eye more than the nose ($p = .038$).

For neutral faces, Mauchly's test indicated that face area did not conform to the assumption of sphericity (Mauchly's W $= .618, \chi^2(5) = 15.3, p = .009$). Therefore, results include Greenhouse-Geisser correction. We found that face area was a significant factor on percent looking time ($F(2.33, 76.8) = 16.3, p < .001, \eta_p^2 = .331$). Post hoc comparisons on region in neutral faces showed that participants spent less time looking at the mouth than any other area of the face (mouth vs. right eye: $p < .001$; mouth vs. left eye: $p < .001$; mouth vs. nose: $p = .009$). In addition, people look at the left eye more than the nose ($p = .047$).

For the sad expression, Mauchly's test indicated that face area did not conform to the assumption of sphericity (Mauchly's W $= .679, \chi^2(5) = 12.3, p = .031$). Therefore, results include Greenhouse-Geisser correction. We found that face area was a significant factor on percent looking time ($F(2.44, 80.5) = 8.97, p < .001, \eta_p^2 = .214$). Post hoc comparisons on region in sad faces showed that participants spent significantly less time looking at the mouth than the eyes (mouth vs. right eye: $p < .001$; mouth vs. left eye: $p = .002$).

For surprise facial expressions, Mauchly's test indicated that face area conformed to the assumption of sphericity (Mauchly's W $= .743, \chi^2(5) = 9.42, p = .094$). We found that face area was a significant factor on percent looking time ($F(3, 99) = 13.3, p < .001, \eta_p^2 = .287$). Post hoc comparisons on region in surprise faces showed that participants looked at the eyes longer than other areas (right eye vs. nose: $p = .009$; right eye vs. mouth: $p < .001$; left eye vs. nose: $p = .001$; left eye vs. mouth: $p < .001$).

Overall the result suggest that the percentage of time looking at certain face areas depends on the emotion category of the stimulus. People spend a significantly shorter amount of time looking at the mouth for the expressions anger, disgust, fear, neutral, and sad. People looked longer at the left eye than the nose for neutral and fear. People looked at the right eye longer for happy faces, and at both eyes longer for surprise.

Predicting emotion category from gaze

Given the differences found across emotions, we then determined which aspects of the gaze were most informative about the emotion being viewed, and which areas of the face were most important. We classified the viewed emotion using BDA and the different GMs for the full trial duration. We first tuned the Gaussian parameter for each map separately using cross validation. Those results are shown in Fig. 5.4. The results show that weighted and unweighted fixation density GMs achieved the best results of approximately 24% 5 fold CV accuracy in predicting the stimuli emotion from the gaze. The analysis will therefore focus on those GMs when appropriate.

We tested to see if these systematic differences were manifested in the overall pattern of fixations or saccades during a segment of the trial as well as over the entire trial. We calculated the GMs for time ranging from the start of the trial to the end of the trial in increments of 200ms, then determined the 5 fold CV accuracy when using BDA to classify the stimuli emotion. We also calculated the GMs for a duration of 400ms from the start of the trial in steps of 200ms. The results are shown in Fig. 5.5. The plot of Fig. 5.5(a) shows that the accuracy in discriminating the emotion category from the GM peaks and stabilizes at about 24% when considering the trial time from the beginning until about 1800ms throughout the end of the 2000ms trial.

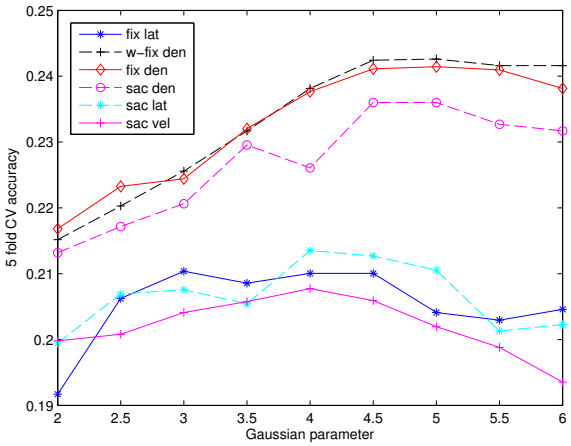
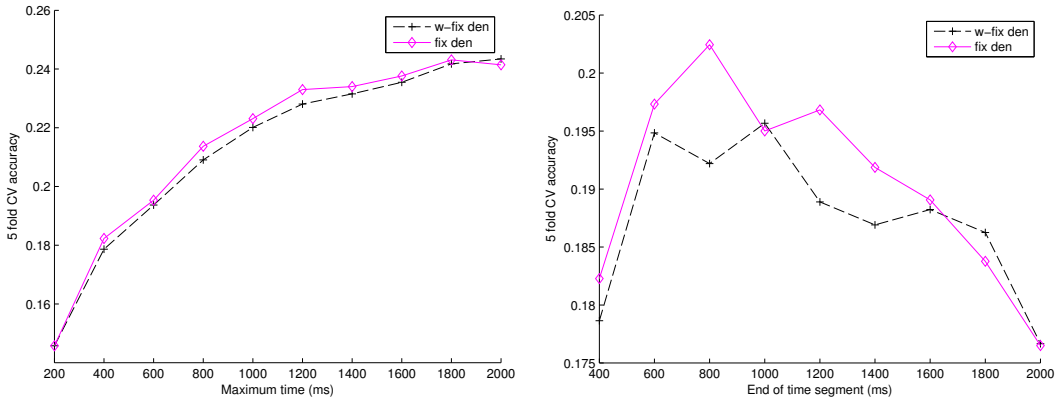


Figure 5.4: 5 fold cross validation accuracy for predicting the stimuli emotion label from the different GMs. The legend entries from top to bottom correspond to fixation latency, weighted fixation density, unweighted fixation density, saccade density, saccade latency, and saccade velocity. The weighted fixation density and unweighted fixation density GMs achieve the best accuracy of 24.3% at $\sigma = 5$ and 24.1% at $\sigma = 5$.

On the other hand, Fig. 5.5(b) shows that there is a peak in discrimination accuracy of about 19.5% for the $200-600$ms time segment until the $600-1000$ms time segment for weighted fixation density. There is a higher peak accuracy of about 20% for the $400-800$ms time segment for unweighted fixation density.

We then visualized the most discriminative dimension for classifying the emotions from the weighted FDM GM. Recall that in a linear classifier, the magnitude of entries of the normal vector for the separating hyperplane gives the importance of the different dimensions. In the case of more than two classes, we no longer find a separating hyperplane, but a one dimensional subspace where we project that data. Similar to the previous case, the magnitude of the entries of that unit vector give the importance of the feature dimensions. Therefore, we plot the absolute value of the subspace in order to visualize the most discriminative regions of the face. We



(a) 0 to x ms

(b) $(x - 400)$ to x ms

Figure 5.5: 5 fold cross validation accuracy for predicting the stimuli emotion over different durations of the trial. Fig. 5.5(a) shows accuracy when considering times from the start of the trial until the end, in increments of 200ms. Fig. 5.5(b) shows accuracy when considering 400ms time segments of the trial, starting with the first 400ms and stepping through the trial in increments of 200ms. We use the same legend conventions as Fig. 5.4.

visualize this overlaid on a neutral face in Fig. 5.6. The figure suggests that the fixation pattern over the entire face is most diagnostic of the emotion category.

Note that the classification performance and the dimensions revealed were determined for classifying all 7 emotions. We acknowledge that there may be differences when classifying pairs of emotions. Therefore, we apply BDA to all pairs using the weighted FDM GM for the full trial, and visualized as before in Fig. 5.7. There are different patterns for the different emotions, although like in the 7 emotion case, the decision classification is influenced by distributed areas of the face.

It may be the case that the the GMs have too much resolution, and the stimuli should be discretized to a small set of AOIs for analysis. We discretize the interior of the face to 4 AOIs including the eyes, nose, and mouth as shown in Fig. 5.3(b). We achieve 5 fold CV accuracy of 19.4%.

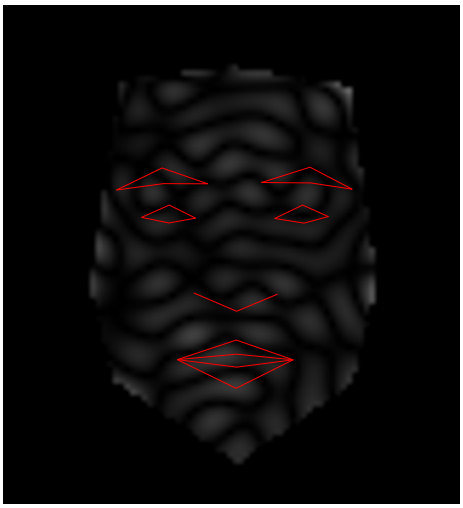


Figure 5.6: Visualization of the absolute value of the most discriminant dimension for classifying all 7 emotion categories using the weighted FDM. The brighter peaks correspond to larger values.

The previous analysis suggests that there are systematic differences in the allocation of attention over the face. Here, we apply different classification models to better understand what may be driving these differences. The success or failure of the models gives insight into the important dimensions of emotion recognition.

We first apply the decision tree classifier. This model suggests that when recognizing emotions, we may first identify some categories while other categories may require further discrimination. This model would explain why some emotions such as happiness and surprise are recognized much easier than anger [28]. We found that the best accuracy was achieved with a single decision node, corresponding to the standard BDA classifier applied earlier. Using additional nodes in the tree classifier only reduced the 5 fold CV classification accuracy to 16.7%.

Figure 5.7: Visualization of the absolute value of the most discriminant dimension for classifying pairs of emotion categories using the weighted FDM. The brighter areas correspond to larger values, and more discriminant areas.

We also explore an alternative that systematic differences in the pattern of viewing different emotional expressions exist in the *sequence* of fixations. We model the fixation sequence as a Markov process having a single random variable taking on 5 possible states corresponding to the eyes, nose and mouth of Fig. 5.3(b), with an additional state for fixations outside those regions. Assuming that The success of this model would suggest that there may be important relational measurements consistently extracted from the face during emotion recognition. We achieve a 5 fold CV accuracy of 21.2% for classifying all 7 emotions when using a MM approach to predict the stimuli emotion label from the fixation sequence.

Finally, we apply the NSS method to predict the emotion label from the fixation sequence. The success of this model suggests that the fixation sequence is dictated by the emotion label of the stimuli. This process may result from differences in the face appearance resulting from representative face deformations, or a top down process directed by categorization. We achieve a 5 fold CV accuracy of 21.4%.

5.2.2 Infant Analysis

We found that neither the looking times nor their transformed versions passed tests of Normality, so we omit the ANOVA analysis of emotion and face area on looking times. Instead, we focus on infant discrimination performance, and a qualitative measure of gaze differences with respect to emotion category. The marginal WFDMs for the emotions tested are shown in Fig. 5.8. In this figure, we have combined the looks on the left and right face during the familiarization trials because the pairs are identical. Unlike the adult marginal maps, there is not a clear distinction between

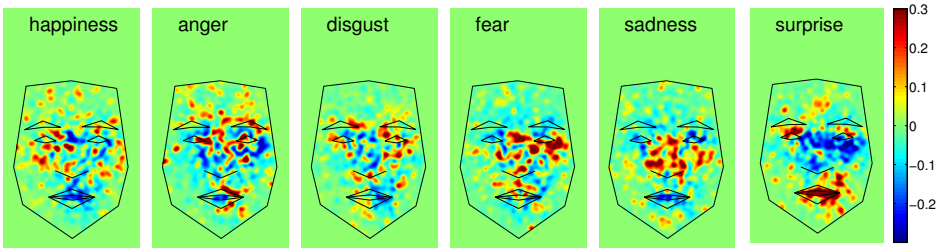


Figure 5.8: Marginal fixation density maps for the different emotions for infants with respect to the overall mean.

the maps, except for the surprise expression where participants clearly look longer at the mouth than the eyes.

To check for evidence of discrimination, we look at the habituation curves and the percentage of novelty preference on the discrimination trials. The percentage novelty preference is given as the total time looking within the novel expression face over the time looking within both the novel and familiarized expression faces. We expect infants that discriminate the expressions to prefer the novel expression face on both trials of each novel expression. Some habituation curves and novelty preference measures are shown for the $6-10$-, $11-13$-, 16- and $20-24$-month-old participants in Figs. 5.9, 5.10, 5.11, and 5.12, respectively.

Since the number of participants and completed trials are limited, it is difficult to assess how these effects generalize. In addition, the order of familiarization and the refamiliarization within a block may have an effect. Therefore, we make some qualitative observations about the data. First, across blocks, there is not a strong novelty preference across emotional categories. In that cases that an infant shows discrimination of two emotional categories, the pattern is not often repeated by other participants in that age group.

In the $6-10$-month-old group, some infants had a novelty preference for sad and disgust faces when familiarized to happy faces. There was also novelty preference for happy when familiarized to surprise faces. In the $11-13$-month-old group, some participants showed novelty preference for surprise, sad, and disgust when familiarized to angry faces. Some showed novelty preference for angry and disgusted faces when familiarized to fearful faces, and novelty preference for happy when familiarized to surprise faces. In the 16-month-old group, some participants showed novelty preference for sad and angry faces when familiarized to disgusted faces. In the final group, there was evidence of preference for happy when familiarized to fear, disgust when familiarized to anger, and disgust when familiarized to surprise. Overall, it is difficult to say whether the age groups are discriminating those categories since the results are inconsistent across participants, and the order of blocks is sure to play a role. Many more participants would be required to determine the generality of these observations.

5.3 Conclusion

Our results suggest that there are some systematic differences in the way adults look at different emotional faces, because we can classify the stimuli emotion based on the fixation density and the sequence of fixations above chance level. However, the gaze alone is insufficient to characterize the emotion label. While some face areas may be more informative for some emotions, gaze patterns suggest that people probably extract holistic face representation across all emotional faces, and make their judgment about the emotion at a higher level of the processing stream.

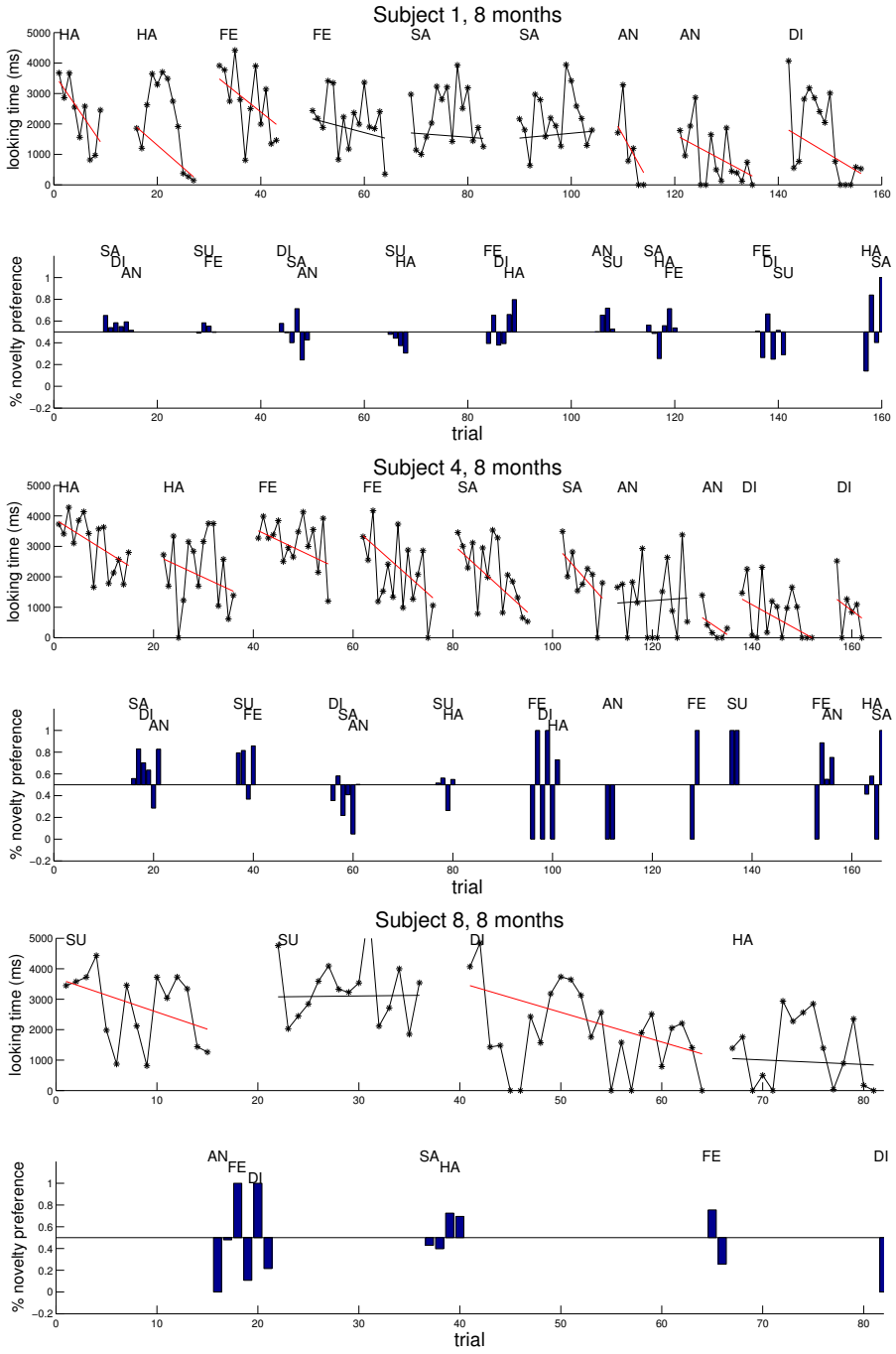


Figure 5.9: Habituation curves and percent novelty preference for infants in the 6−10-month-old group. Red lines over the habituation curves denote a 30% drop in looking time.

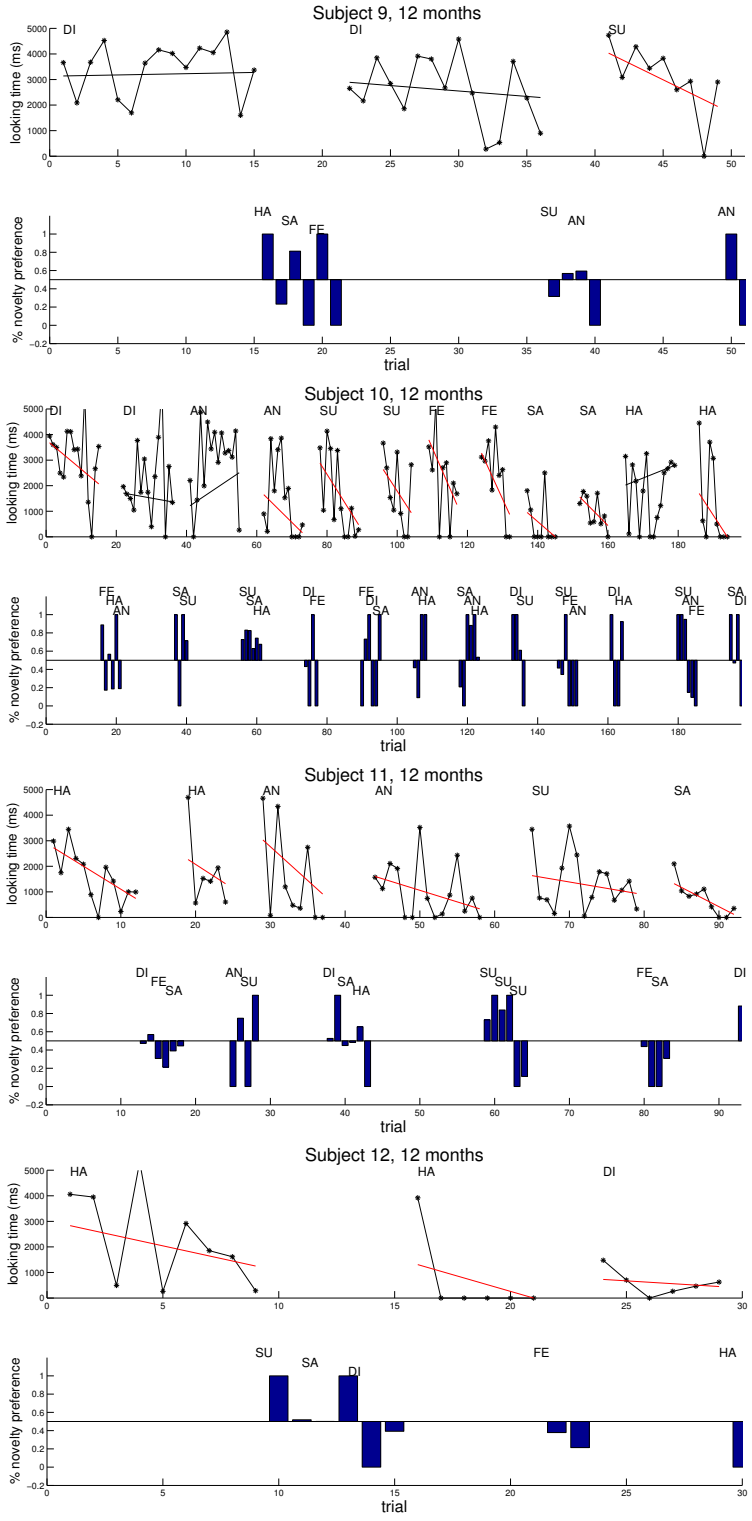


Figure 5.10: Habituation curves and percent novelty preference for infants in the $11-13$-month-old group. Red lines over the habituation curves denote a 30% drop in looking time.

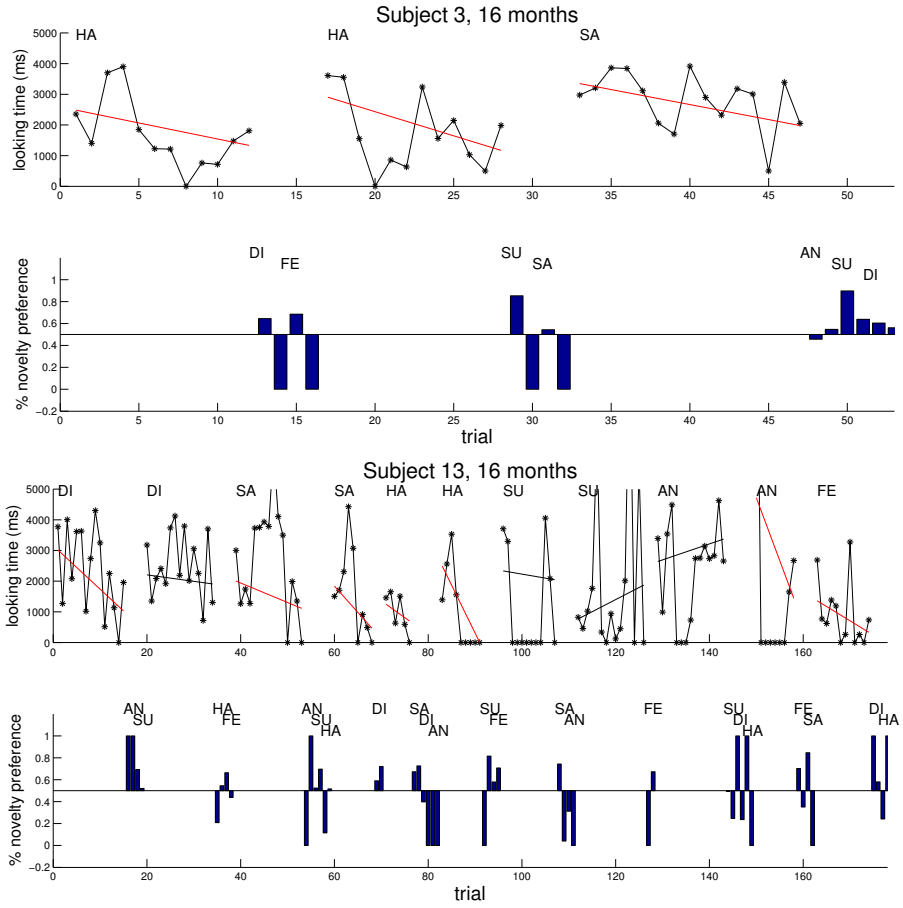


Figure 5.11: Habituation curves and percent novelty preference for infants in the 16-month-old group. Red lines over the habituation curves denote a 30% drop in looking time.

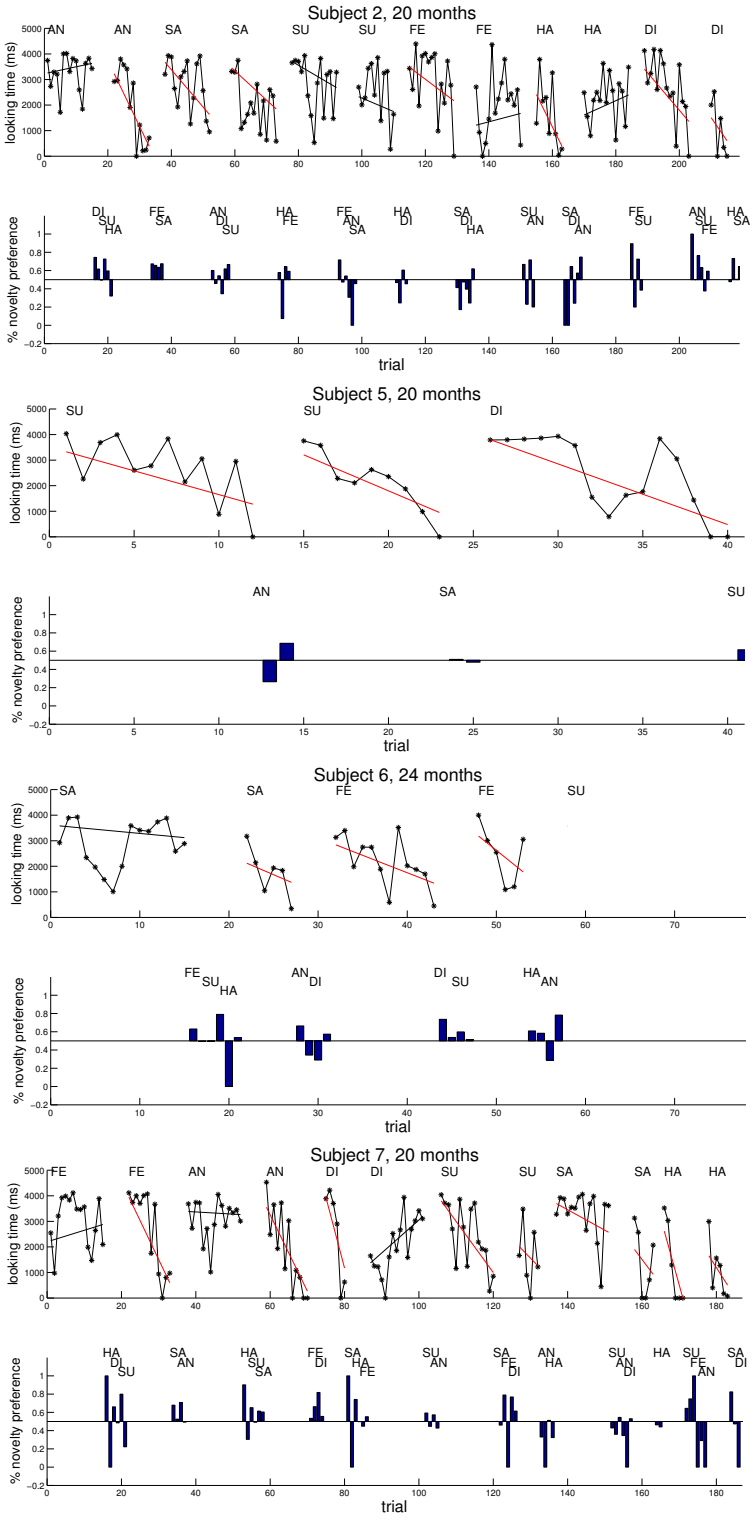


Figure 5.12: Habituation curves and percent novelty preference for infants in the $20-24$-month-old group. Red lines over the habituation curves denote a 30% drop in looking time.

Infants are shown to look more uniformly at the face across different emotions, except for the surprise expression where they look longer at the mouth. However, we show some evidence that infants from as early as 8 months can discriminate emotional expression categories across changed identities. However, more data is required to make definitive claims since the sample size within age groups is very small.

CHAPTER 6

CONTRIBUTIONS, CONCLUSIONS, AND FUTURE DIRECTIONS

We have made several contributions regarding deformable shape detection, the identification of relevant eye tracking variables, and the dimensions for recognizing emotional facial expressions. In this chapter we summarize our major contributions and conclusions, and discuss directions for future work.

6.1 Contributions and Conclusions

Our manifold learning shape detection work has shown that it is possible to learn the function relating an object image to the shape, so that detailed face shape can be detected in one iteration. The method is flexible enough to handle different identities and expressions, and can work in very low resolution settings. We showed that while other regression approaches can give similar results in high resolution settings, the local feature based regression methods do not work as well in the very low resolution case. Our other shape detection algorithm based on probabilistic graphical models combines the positive aspects of local feature detection and global shape modeling in order to provide a precise and dense shape detection. Previously, graphical approaches have only detected a small set of salient points, whereas we detect many salient and

non-salient landmarks. In addition, we showed that our approach generalizes across databases better than global appearance modeling algorithms like AAM.

Our methodology for automatically identifying the variables relevant to categorization has found a small set of variables that predict if participants learned visual categories. Previously, ad hoc approaches were used to determine which variables to analyze in a study - possibly ignoring important aspects of the data. Instead, our method determines which aspects of gaze are informative from a large pool of possible variables. In addition, the results validate the use of several variables such as proportion fixation and fixation latency that have been used in previous categorization studies. More interesting is the finding that the variables corresponding to the overall pattern of infant looking allow us to predict if infants learned the category. This suggests that infants are not directing their attention randomly.

In our emotion perception work we have shown that although the gaze does differ slightly when viewing different emotional expressions during a categorization task, the gaze alone may be insufficient to discriminate the emotion category. The methodology presented also demonstrates how to compare gaze across non-schematic faces, and how to determine which aspects of the gaze are most informative. Specifically, we found that fixation density was more informative that the saccade and latency information. In the infant experiments we found evidence that infants probably can discriminate between even the difficult emotional expression categories such as disgust and anger. We also found that qualitatively, infants look at the same internal areas of the face as adults.

6.2 Future Work

We acknowledge that this research has built on several earlier studies, and we offer some ideas and directions for the extension and continuation of our research. In the case of the shape detection by manifold learning, the main difficulties are interpolating the shape of test samples from different databases, and detecting subtle shape deformations such as single eye blinks that are not well represented in the training set. Given that the method relies on a regression model, an image feature space, and a shape model, the obvious direction for progress is to improve those components of the approach. For example, we can use a different loss function for the regressor, or a different feature space for the image space.

In our unpublished work, we have explored detecting the shape of components of the face such as the eyes and mouth separately. We found that this can yield precise detection of those components since the shape model must only account for the deformation of local shape regions, but requires that the components are localized precisely in the test image. We also found that the detection performance suffers from poor normalization of the faces before detection, and by iteratively optimizing the object location and shape, we can improve accuracy. The difficult with this approach is that we must establish some criterion for how well the face is localized. We have tried using probabilistic texture models as in AAM, but occluded faces yield a low probability even when they are precisely localized. One promising direction for future research is to develop a robust criterion for proper localization to assist subsequent detection of the face shape, or its parts.

Our probabilistic graphical model can benefit from a more flexible graph structure, and a faster inference procedure. The potential functions describing the expected

relative positions of landmarks assume a Gaussian form, which may be inappropriate for some landmark pairs. For example, the distance between the top and bottom lip may exhibit a bimodal distribution corresponding to an open or closed mouth, which is not modeled well by a single Gaussian distribution. In addition, the inference procedure is too slow to detect a dense set of points in real time.

In our variable selection for categorization study we identified some variables that predict category learning, but it is not certain whether these variables are generally important to categorization or specific to this experiment. For example, would we still find that the proportion of looking at the relevant area is predictive of category learning in adults if no supervision took place? Even more fundamental of a contribution would be a procedure for comparing infants to adults. As mentioned in chapter 4, the infants and adults have different goals during the categorization task since unlike infants, adults are explicitly asked to learn and identify a category. Equating the goals of adults and infants during the categorization task may make comparison of gaze across infant and adult populations more meaningful.

The results of our emotion perception study support a model where participants encode a common or partly overlapping set of dimensions when categorizing emotional faces. The next step is to identify the dimensions that would lead to the pattern of perception that we find in experiments. For example, if we expect shape information such as distances between face fiducials to be important, which of the many relations in the face are used? One approach would be to conduct another study where all but a few possible dimensions are masked from the face, as in the bubbles study, and find the dimensions that allow recognition. The problem with that approach is that participants may rely on an analytical local feature processing in order to categorize

the expression if only particular features are visible. Therefore, we could not be certain that the dimensions found to be important using this paradigm were the ones used by participants in a more natural setting. A better approach would be to use natural looking faces where particular dimensions were altered, and test participants to see how and if their perception of the faces changed. Such a methodology has been used successfully in the past [86], but only tested a few dimensions and emotions. If we are to test a wide variety of dimensions and emotions, it would be necessary to carefully determine which dimensions to alter for the subsequent behavioral study since testing every combination would be impractical.

We began exploring this problem of identifying the important dimensions by attempting to determine which relational measures of the face stimuli from the adult emotion study yield the pattern of perception that we see in the adult participants. Specifically, we have about 180 expressive face stimuli that were categorized by 34 subjects who make mistakes about 23% of the time. We hypothesize that some of these mistakes result from similarity in the value of one or more of the important relational dimensions between the true and perceived emotional face category. Therefore, by considering all the relational measure values and the pattern of emotion perception for all of the tested stimuli, we may be able to find the dimensions that would predict the pattern of perception that we found in the data. The next step would be to validate those dimensions, either through another behavioral study or by cross validation.

In the case of infants, the main shortcoming of this work is the lack data. We cannot say with confidence that infants are discriminating emotional faces since we only have about 5 participants in each age group, and most participants did not

complete the study. We would like to collect sufficient data in order to determine if there are statistically significant changes in gaze over the different age groups, and to make more specific conclusions about the ability of those age groups to discriminate the different emotional facial expressions. Since our design takes about 20 minutes for an infant to complete, one direction of future research would be to develop a paradigm which tests discrimination of the 6 emotions in half the time. Since habituation and forced preferential looking take a significant amount of type, it may be necessary to develop an alternative paradigm for assessing discrimination in infants.

BIBLIOGRAPHY

[1] Floyd Henry Allport. *Social Psychology*. Boston: Houghton Mifflin Company, 1924.

[2] Dima Amso and Scott P. Johnson. Selection and inhibition in infancy: evidence from the spatial negative priming paradigm. *Cognition*, 95(2):B27–B36, 2005.

[3] Dima Amso and Scott P. Johnson. Learning by selection: Visual search and object perception in young infants. *Developmental Psychology*, 42(6):1236–1245, 2006.

[4] Dima Amso and Scott P. Johnson. Development of visual selection in 3- to 9-month-olds: Evidence from saccades to previously ignored locations. *Infancy*, 13:675–686, 2008.

[5] M.F. Augusteijn and T.L. Skufca. Identification of human faces through texture-based feature recognition and neural network technology. In *Neural Networks, 1993., IEEE International Conference on*, pages 392 –398 vol.1, 1993.

[6] Robert J. Baron. Mechanisms of human facial recognition. *International Journal of Man-Machine Studies*, 15(2):137 – 178, 1981.

[7] Harry G Barrow and J Martin Tenenbaum. Recovering intrinsic scene characteristics from images. In *Computer Vision Systems*. Academic Press, 1978.

[8] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2835–2404, 2000.

[9] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, September 1999.

[10] C. A. Best, C. W. Robinson, and V. M. Sloutsky. The effect of labels on visual attention: An eye tracking study. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1846–1851, 2010.

[11] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999.

[12] Sven Bolte and Fritz Poustka. The recognition of facial affect in autistic and schizophrenic subjects and their first-degree relatives. *Psychological Medicine*, 33(5):907–915, 2003.

[13] Gary R. Bradski and Adrian Kaehler. *Learning OpenCV - computer vision with the OpenCV library: software that sees.* O'Reilly Media, 2008.

[14] R A Brooks, R Greiner, and T O Binford. The acronym model-based vision system. In *IJCAI*, pages 105–113, 1979.

[15] Rodney A. Brooks. Model-based three dimensional interpretations of two dimensional images. In *IJCAI*, pages 619–624, 1981.

[16] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[17] Rose F. Caron, Albert J. Caron, and Rose S. Myers. Abstraction of invariant face expressions in infancy. *Child Development*, 53(4):1008–1015, 1982.

[18] D. Chai and K.N. Ngan. Locating facial region of a head-and-shoulders color image. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 124 –129, apr 1998.

[19] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[20] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *In Proc. European Conference on Computer Vision*, pages 484–498, 1998.

[21] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[22] M. Corbetta, E. Akbudak, T. Conturo, A. Snyder, J. Ollinger, H. Drury, M. Linenweber, S. Petersen, M. Raichle, D. Vanessen, and G. L. Shulman. A common network of functional areas for attention and eye movements. *Neuron*, 21(4):761–773, Oct 1998.

[23] D. Cristinacce and T. F. Cootes. Boosted regression active shape models. *In Proc. British Machine Vision Conference*, pages 880–889, 2007.

[24] Charles Darwin. *The Expression of the Emotions in Man and Animals.* The University of Chicago Press, 1965.

[25] G. B. Duchenne de Boulogne. *The Mechanism of Human Facial Expression,*. Cambridge University Press, 1990.

[26] L. Ding and A.M. Martinez. Precise detailed detection of faces and facial features. *In Proc. IEEE Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[27] L. Ding and A.M. Martinez. Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 2022–2038, March 2010.

[28] Shichuan Du and Aleix M. Martinez. The resolution of facial expressions of emotion. *Journal of Vision*, 11(13):24,1–13, 2011.

[29] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edition, 2001.

[30] Hedwig Eisenbarth and Georg W. Alpers. Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion*, 11(4):860–865, 2011.

[31] P. Ekman. *Emotion in the Human Face*. Cambridge University Press, 1982.

[32] Paul Ekman and Wallace Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[33] Paul Ekman and Wallace V. Friesen. *Pictures of Facial Affect*. Consulting Psychologists Press, 1976.

[34] Ralf Engbert and Reinhold Kliegl. Microsaccades uncover the orientation of covert attention. *Vision Research*, 43(9):1035 – 1045, 2003.

[35] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *BMVC*, 2006.

[36] M. Everingham and A. Zisserman. Regression and classification approaches to eye localization in face images. *In Proc. IEEE Automatic Face and Gesture Recognition*, pages 441–446, April 2006.

[37] Terje Falck-Ytter, Gustaf Gredebäck, and Claes von Hofsten. Infants predict other people's action goals. *Nature Neuroscience*, 9(7):878–879, 2006.

[38] M. J. Farah, K. D. Wilson, M. Drain, and J. N. Tanaka. What is "special" about face perception? *Psychological review*, 105(3):482–498, July 1998.

[39] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[40] Tiffany M. Field, Robert Woodson, Reena Greenberg, and Debra Cohen. Discrimination and imitation of facial expressions by neonates. *Science*, 218(8):179–181, 1982.

[41] N H Frijda. Emotion and recognition of emotion. In *Paper presented at the Third Symposium on Feelings and Emotions, Loyola University, Chicago, October 10-12*, 1968.

[42] N H Frijda and E Philipszoon. Dimensions of recognition of emotion. *Journal of Abnormal Social Psychology*, 66:45–51, 1963.

[43] Seymour Geisser. Discrimination, allocatory and separatory, linear aspects. *Classification and Clustering*, J. Van Ryzin (Ed.):301–330, 1977.

[44] Stuart Geman and Donald Geman. Readings in computer vision: issues, problems, principles, and paradigms. chapter Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, pages 564–584. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987.

[45] F. Gosselin and P. G. Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research*, 41(17):2261–2271, August 2001.

[46] Alex Gray and Edward Moore. Very fast multivariate kernel density estimation using via computational geometry. In *Proceedings, Joint Stat*, 2003.

[47] Leon Gu and Takeo Kanade. A generative shape regularization model for robust face alignment. In *ECCV (1)*, volume 5302 of *Lecture Notes in Computer Science*, pages 413–426. Springer, 2008.

[48] Leon Gu and Takeo Kanade. A generative shape regularization model for robust face alignment. *In Proc. European Conference on Computer Vision*, pages 413–426, 2008.

[49] Lie Gu and Takeo Kanade. 3d alignment of face in a single image. pages 1305–1312. CVPR, 2006.

[50] Kun Guo. Holistic gaze strategy to categorize facial expression of varying intensities. *PLoS ONE*, 7(8):e42585, 08 2012.

[51] O. C. Hamsici and A. M. Martinez. Active appearance models with rotation invariant kernels. *In Proc. IEEE International Conference on Computer Vision*, pages 1003–1009, 2009.

[52] O. C. Hamsici and A. M. Martinez. Rotation invariant kernels and their application to shape analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(11):1985–1999, 2009.

[53] Onur C. Hamsici and Aleix M. Martinez. Bayes optimality in linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:647–657, 2008.

[54] A. R. Hanson and E. M. Riseman. VISIONS: A computer system for interpreting scenes. In *Computer Vision Systems*. Academic Press, 1978.

[55] L.D. Harmon, M.K. Khan, Richard Lasch, and P.F. Ramig. Machine identification of human faces. *Pattern Recognition*, 13(2):97 – 110, 1981.

[56] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001.

[57] A H Hastorf, C E Osgood, and H Ono. The semantics of facial expressions and the prediction of the meanings of stereoscopically fused facial expressions. *Scandinavian Journal of Psychology*, 7(3):179–188, 1966.

[58] Nadia Hernandez, Aude Metzger, Rmy Magn, Frdrique Bonnet-Brilhault, Sylvie Roux, Catherine Barthelemy, and Jolle Martineau. Exploration of core features of a human face by healthy and autistic adults analyzed by visual scanning. *Neuropsychologia*, 47(4):1004–1012, 2009.

[59] A. E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58:54–59, 1962.

[60] Ming hsuan Yang and Narendra Ahuja. Gaussian mixture model for human skin color and its applications in image and video databases. In *Proceedings of SPIE: Storage and Retrieval for Image and Video Databases VII*, pages 458–466, 1999.

[61] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *University of Massachusetts, Amherst, Technical Report 07-49, October*, 2007.

[62] S. P. Johnson, D. Amso, and J. A. Slemmer. Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences*, 100(18):10568–10573, 2003.

[63] S. P. Johnson, J. Davidow, C. Hall-Haro, and M. C. Frank. Development of perceptual completion originates in information acquisition. *Developmental Psychology*, 44(5):1214–1224, 2008.

[64] S. P. Johnson, J. A. Slemmer, and D.. Amso. Where infants look determines how they see: Eye movements and object perception performance in 3-month-olds. *Infancy*, 6(2):185–201, 2004.

[65] Michael Kass, Andrew Witkins, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331, 1988.

[66] Heidi Kloos and Vladimir M. Sloutsky. What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, 137(1):52–72, 2008.

[67] Constantine Kotropoulos and Ioannis Pitas. Rule-based face detection in frontal views. In *in Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 97), vol. IV*, pages 2537–2540, 1997.

[68] T. K. Leung, M. C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *ICCV '95: Proceedings of the Fifth International Conference on Computer Vision*, pages 637+, Washington, DC, USA, 1995. IEEE Computer Society.

[69] Peng Li and S.J.D. Prince. Joint and implicit registration for face recognition. volume 0, pages 1510–1517, Miami, FL, USA, 2009. IEEE Computer Society.

[70] Lin Liang, Fang Wen, Xiaoou Tang, and Y Xu. An integrated model for accurate shape alignment. *In Proc. European Conference on Computer Vision*, pages 333–346, 2006.

[71] Xiaoming Liu. Discriminative face alignment. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 99(2):1941–1954, 2009.

[72] R. Malladi, J.A. Sethian, and B.C. Vemuri. Shape modeling with front propagation: A level set approach. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17:158–175, 1995.

[73] A. M. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:748–763, 2002.

[74] A. M. Martinez and R. Benavente. The AR face database. *CVC Technical Report No. 24*, 1998.

[75] A. M. Martinez and Manli Zhu. Where are linear feature extraction methods applicable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12):1934–1944, 2005.

[76] Bob McMurray and Richard N. Aslin. Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy*, 6(2):203–229, 2004.

[77] K. Messer, J. Matas, J. Kittler, and K. Jonsson. Xm2vtsdb: The extended m2vts database. In *In Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[78] K. Messer, J. Matas, J. Kittler, J. Lüttin, and G. Maitre. Xm2vtsdb: The extended m2vts database. *In Proc. International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[79] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher discriminant analysis with kernels. In *Proc. IEEE Neural Networks for Signal Processing Workshop*, pages 41–48, 1999.

[80] Greg Mori, Serge Belongie, and Jitendra Malik. Efficient shape matching using shape contexts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27:517–530, 2005.

[81] Tsuyoshi Moriyama, Takeo Kanade, Jing Xiao, and Jeffrey F. Cohn. Meticulously detailed eye region model and its application to analysis of facial images. *PAMI*, 28:2006, 2006.

[82] Kevin Murphy. Kalman filter toolbox for matlab, 2004.

[83] C A Nelson. The perception and recognition of facial expressions in infancy. In T. M. Field & N. A. Fox, editor, *Social perception in infants*, pages 101–125. Norwood, NJ: Ablex, 1985.

[84] Charles A. Nelson. The recognition of facial expressions in the first two years of life: Mechanisms of development. *Child Development*, 58(4):pp. 889–909, 1987.

[85] Charles A. Nelson and Kim G. Dolgin. The generalized discrimination of facial expressions by seven-month-old infants. *Child Development*, 56(1):pp. 58–61, 1985.

[86] Donald Neth and Aleix M. Martinez. A computational shape-based model of anger and sadness justifies a configural representation of faces. *Vision Research*, 50(17):1693 – 1711, 2010.

[87] C E Osgood. Dimensionality of the semantic space for communication via facial expressions. *Scandinavian Journal of Psychology*, 7:1–30, 1966.

[88] K A Pelphrey, N J Sasson, J S Reznick, G Paul, B D Goldman, and J Piven. Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders*, 32(4):249–261, 2002.

[89] R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, August 2005.

[90] R Plutchik. *The emotions: facts, theories, and a new model*. New York: Random House, 1962.

[91] Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[92] P. C. Quinn, P. D. Eimas, and S. L. Rosenkrantz. Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22(4):463–475, 1993.

[93] Paul C. Quinn, Matthew M. Doran, Jason E. Reiss, and James E. Hoffman. Time course of visual attention in infant categorization of cats versus dogs: evidence for a head bias as revealed through eye tracking. *Child development*, 80(1):151–161, 2009.

[94] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[95] B. Rehder and A. B. Hoffman. Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51:1–41, 2005.

[96] M. Rutherford and Ashley Towns. Scan path differences and similarities during emotion perception in those with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 38:1371–1381, 2008.

[97] J M Douglas S Vassallo, S L Cooper. Visual scanning in the recognition of facial affect: is there an observer sex difference? *Journal of Vision*, 9(3):1–10, 2009.

[98] T. Sakai, M. Nagao, and S. Fujibayashi. Line extraction and pattern detection in a photograph. *Pattern Recognition*, 1(3):233 – 248, 1969.

[99] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *ETRA '00: Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78, New York, NY, USA, 2000.

[100] Ashok Samal and Prasana A. Iyengar. Human face detection using silhouettes. *International Journal of Pattern Recognition and Artificial Intelligence*, 09(06):845–867, 1995.

[101] Brian Scassellati. Eye finding via face detection for a foveated active vision system. In Jack Mostow and Chuck Rich, editors, *AAAI/IAAI*, pages 969–976. AAAI Press / The MIT Press, 1998.

[102] Mark Schmidt. L1general - matlab code for solving l1-regularization problems, 2011.

[103] H. Scholsberg. A scale for the judgment of facial expressions. *Journal of Experimental Psychology*, 29(6):497–510, 1941.

[104] Gail M. Schwartz, Carroll E. Izard, and Susan E. Ansul. The 5-month-old's ability to discriminate facial expressions of emotion. *Infant Behavior and Development*, 8(1):65–77, 1985.

[105] G.A.F. Seber and C.J. Wild. *Nonlinear Regression*. Wiley-Interscience, September 2003.

[106] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(3):411–426, March 2007.

[107] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, illustrated edition edition, jun 2004.

[108] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *CVPR 2007*, June 2007.

[109] Saad Ahmed Sirohey, Masooda Begum, Iftikhar A. Sirohey, and Zarina Sirohey. Human face segmentation and identification, 1993.

[110] Daniel T Smith, Chris Rorden, and Stephen R Jackson. Exogenous orienting of attention depends upon the ability to execute eye movements. *Current Biology*, 14(9):792 – 795, 2004.

[111] Fraser W. Smith and Philippe G. Schyns. Smile through your fear and sadness: Transmitting and identifying facial expression signals over a range of viewing distances. *Psychological Science*, 20(10):1202–1208, 2009.

[112] Marie L. Smith, Garrison W. Cottrell, Frederic Gosselin, and Philippe G. Schyns. Transmitting and decoding facial expressions. *Psychological Science*, 16(3):184–189, March 2005.

[113] Alex J. Smola and Bernhard Schlkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

[114] Dave M Stampe. Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavioral Research Methods, Instruments, & Computers*, 25(2):137–142, 1993.

[115] M. B. Stegmann and D. D. Gomez. A brief introduction to statistical shape analysis. Informatics and Mathematical Modelling, Technical University of Denmark, DTU, mar 2002.

[116] D N Stern. Mother and infant at play: The dyadic interaction involving facial, vocal, and gaze behavior. In M. Lewis & L. Rosenblum, editor, *The effect of the infant on its caregiver*, pages 187–213. New York: Wiley, 1974.

[117] T.M. Strat and M.A. Fischler. Context-based vision: recognizing objects using information from both 2d and 3d imagery. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13:1050–1065, 1991.

[118] Raphael Sznitman and Bruno Jedynak. Active testing for face detection and localization. *PAMI*, 32:1914–1920, 2010.

[119] Y.-I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97 –115, feb 2001.

[120] S S Tomkins and R McCarter. What and where are the primary affects? some evidence for a theory. *Perceptual and Motor Skills*, 18(1):119–1589, 1964.

[121] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, November 1999.

[122] I Vine. The role of facial-visual signaling in early social development. In M. Von Cranach and I. Vine, editors, *Social communication and movement: Studies of interaction and expression in Man and Chimpanzee*, pages 195–298. London: Academic Press, 1973.

[123] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. *In Proc. IEEE Computer Vision and Pattern Recognition*, pages I–511–I–518, 2001.

[124] Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[125] R S Woodworth. *Experimental psychology*. New York: Henry Holt, 1938.

[126] Guangzheng Yang and Thomas S Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53 – 63, 1994.

[127] Ming-Hsuan Yang, D.J. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(1):34 –58, jan 2002.

[128] D. You, O. Hamsici, and A. M. Martinez. Kernel optimization in discriminant analysis. *PAMI*, 33, 2011.

[129] Andrew W. Young, Duncan Rowland, Andrew J. Calder, Nancy L. Etcoff, Anil Seth, and David I. Perrett. Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition*, 63(3):271–313, July 1997.

[130] Gail Young-Browne, Howard M. Rosenfeld, and Frances Degen Horowitz. Infant discrimination of facial expressions. *Child Development*, 48(2):pp. 555–562, 1977.

[131] Jingdan Zhang, Shaohua Kevin Zhou, Dorin Comaniciu, and Leonard McMillan. Conditional density learning via regression with application to deformable shape segmentation. *In Proc. IEEE Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[132] F. Zhou, F. De la Torre, and J.F. Cohn. Unsupervised discovery of facial events. In *CVPR*, 2010.

[133] Shaohua Zhou and Dorin Comaniciu. Shape regression machine. *Information Processing in Medical Imaging*, pages 13–25, 2007.

[134] Shaohua K. Zhou. Shape regression machine and efficient segmentation of left ventricle endocardium from 2D b-mode echocardiogram. *Medical Image Analysis*, apr 2010.

[135] Yi Zhou, Lie Gu, and Hong-Jiang Zhang. Bayesian tangent shape model: estimating shape and pose parameters via bayesian inference. *In Proc. IEEE Computer Vision and Pattern Recognition*, pages I–109–I–116, 2003.